

GENERALIZED EMPIRICAL BAYES:
THEORY, METHODOLOGY, AND APPLICATIONS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fullfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Douglas M. Fletcher II
Diploma Date May 2019

Examining Committee Members:

Dr. Subhadeep Mukhopadhyay, Dissertation Advisor, Department of
Statistical Science

Dr. Alan Izenman, Dissertation Examination Chair, Department of
Statistical Science

Dr. William Wei, Department of Statistical Science

Dr. (LTC) Randal Hickman, Army Cyber Institute

Dr. Iyad Obeid, Department of Electrical and Computer Engineering

©

by

Douglas M. Fletcher II

December, 2018

All Rights Reserved

ABSTRACT

GENERALIZED EMPIRICAL BAYES: THEORY, METHODOLOGY, AND APPLICATIONS

Douglas M. Fletcher

DOCTOR OF PHILOSOPHY

Temple University, December 2018

Dr. Subhadeep Mukhopadhyay, Dissertation Advisor

The two key issues of modern Bayesian statistics are: (i) establishing a principled approach for *distilling* a statistical prior distribution 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 thesis, we propose generalized empirical Bayes as a new framework for exploring these fundamental questions along with a wide range of applications spanning fields as diverse as clinical trials, metrology, insurance, medicine, and ecology. Our research marks a significant step towards bridging the “gap” between Bayesian and frequentist schools of thought that has plagued statisticians for over 250 years.

Chapters 1 and 2—based on Mukhopadhyay and Fletcher (2018b)—introduces the core theory and methods of our proposed generalized empirical Bayes (gEB) framework that solves a long-standing puzzle of modern Bayes, originally posed by Herbert Robbins (1980). One of the main contributions of this research is to introduce and study a new class of nonparametric priors $DS(G, m)$ that allows exploratory Bayesian modeling. However, at a practical level, 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. Connections with other Bayesian cultures are also presented in the chapter.

Chapter 3 deals with the topic of measurement uncertainty from a new angle by introducing the foundation of nonparametric meta-analysis. We have applied the proposed methodology to real data examples from astronomy, physics, and medical disciplines. Chapter 4 discusses some further extensions and application of our theory to distributed big data modeling and the missing species problem. The dissertation concludes by highlighting two important areas of future work: a full Bayesian implementation workflow and potential applications in cybersecurity.

DEDICATION

To my wonderful children, Maisie and Milo. While this dissertation may be some evidence that Daddy is as smart as Mommy, we all know the truth...

ACKNOWLEDGMENTS

First, I want to thank my advisor Dr. Subhadeep ‘Deep’ Mukhopadhyay. Your positive energy, forward-thinking, and desire for excellence made this experience both profoundly meaningful and extremely fun. Our path together saw highs and lows, but your commitment to quality and ground-breaking research remained a persistent beacon throughout the entire journey. Thank you for pushing me out of my comfort zone and forcing me to find and answer the ‘true’ questions. We have achieved a lot during this journey, and I am excited for our future efforts.

In addition to my advisor, I want to express my appreciation for my examining committee members: Professor Alan Isenman, Professor William Wei, Professor Iyad Obeid, and Lieutenant Colonel (Dr.) Randal Hickman. Thank you for providing great support, insightful feedback and challenging questions, all of which helped me excel.

I would also like to thank Dr. Richard Heiberger. Not only did you help me realize the power and elegance of good R code, you helped me get ‘re-calibrated’ as a student after being away for quite a while. Thank you for your constructive input to my research and dissertation. I truly value your mentorship and passion.

Finally, I want to thank my amazing wife Hilary. Throughout this endeavor, you have been my source of support, motivation, and love. Words cannot express how much you mean to me, nor how much I treasure all you do for our family. In terms of this dissertation, I am the g and you are the $d[G]$. Together, we make an outstanding $\hat{\pi}$. I love you!

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF TABLES	ix
LIST OF FIGURES	xi
1 INTRODUCTION AND FOUNDATIONS	1
1.1 Introduction	1
1.2 Priors	4
1.2.1 Traditional Approaches	5
1.2.2 Empirical Approaches	7
1.2.3 Which Philosophy is Best?	9
1.3 A Third Culture	9
1.4 Summary of Contributions	11
2 A NEW CLASS OF PRIOR DISTRIBUTION MODELS	13
2.1 The Model	13
2.1.1 New Family of Prior Densities	13
2.1.2 Exploratory Diagnostics and U-Function	18
2.2 Estimation Method	23
2.2.1 Theory	23
2.2.2 Algorithm	26
2.2.3 Maximum Entropy Representation	28
2.2.4 Results	30
2.3 Inference	32
2.3.1 MacroInference	32
2.3.2 Learning From Uncertain Data	36
2.3.3 MicroInference	39
2.3.4 Poisson Smoothing: The Two Cultures	45
2.4 Incorporating Covariates	47
2.4.1 The Norberg Example	48
2.4.2 The Galaxy Example	50
2.5 Connections	53
2.5.1 Robust Bayesian Methods	53
2.5.2 Empirical Bayes Methods	54

2.5.3	Dirichlet-Process-based Approaches	56
2.5.4	Weakly Informative Priors	57
2.6	Software	58
2.7	Summary	58
3	UNCERTAINTY MODELING	60
3.1	Dealing with Uncertain Data	61
3.1.1	Tier 1	62
3.1.2	Tier 2	63
3.1.3	Tier 3	63
3.1.4	Tier 4	67
3.2	Real Data Applications of Tiered Analysis	70
3.2.1	Hubble Constant H_0	70
3.2.2	Newton's Gravitational Constant G_N	74
3.2.3	Lithium Abundance ${}^7\text{Li}$	79
3.3	Summary	80
4	ADDITIONAL APPLICATIONS	83
4.1	Inference from Distributed Data	84
4.1.1	V-Data	88
4.1.2	Cheese Data	91
4.1.3	Key Observations	95
4.2	The Missing Species problem	96
4.2.1	g -Modeling	99
4.2.2	f -Modeling	107
4.2.3	Key Observations	118
4.3	Summary	118
5	CONCLUSION AND FUTURE WORK	120
5.1	Conclusion	120
5.2	Moving toward Full Bayesian Analysis	121
5.3	Modeling Attack Types for Improved Cybersecurity	124
	BIBLIOGRAPHY	126
	APPENDIX	137

LIST OF TABLES

2.1	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.	24
2.2	Two group partitions of the rat tumor studies based on K-means clustering on the posterior mode predictions (see Section 2.3.3 and Fig. 2.10(c)).	36
2.3	Measurements (sorted) along with their uncertainty from different laboratories in arsenic data.	37
2.4	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.	46
2.5	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™ i5-7200 CPU @ 2.50GHz with Windows 10. DPpackage uses C++ compiler to speed-up, while ours is a prototype version implemented in R.	56
3.1	Quantile values for the Hubble data set.	73
3.2	Quantile values for the Big G data set.	77
3.3	Comparison of original modes identified from $\hat{f}(z)$ to the refined clusters identified by applying k-means to the $ z_{ij} $	77
4.1	Comparison of parameters for the cheese data derived from a pooled Bayesian regression and a distributed embarrassingly parallel framework.	93
4.2	The butterfly data, where the count of species seen y times during Corbet’s two-years of trapping.	97
4.3	Comparison of the bias and variance between truncation points $y_0 = 9$ and $y_0 = 19$. For the Bias range, y_+ is the sum of all the y values up to the truncation point y_0	112
4.4	Comparison of results using the DS-distribution algorithm with binomial smoothing against other f and g modeling techniques.	114

4.5	The frequencies for the unique word types that Shakespeare used in his works, y_x is the total number of word types y used x times (Efron and Thisted, 1976). Not shown are 846 word types that appear more than 100 times.	115
-----	---	-----

LIST OF FIGURES

2.1	Proposed model-building approach for generalized empirical Bayes. . .	14
2.2	LP-polynomials $T_j(\theta; G_{\alpha,\beta})$ for <code>family= "beta"</code> with the following (α, β) choices: (a) Jeffrey's prior ($\alpha = \beta = 0.5$), (b) uniform prior ($\alpha = \beta = 1$), and (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.	16
2.3	An illustration of how m allows the DS(G, m) prior flexibility to deviate from $g(\theta)$, should the data indicate $g(\theta)$ requires correction.	18
2.4	Graphical diagnostic tool: U-functions for (a) rat tumor data; (b) terbinafine and ulcer data; and (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.	20
2.5	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.	30
2.6	Plots of penalized uncertainty quantification with respect to m for (a) rat tumor data, (b) shipyard data, and (c) surgical node data. The peak bend in each plot indicates the optimal m for each data set. . .	31
2.7	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.	33
2.8	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).	35
2.9	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).	37

2.10	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($\hat{\alpha} = 2.3, \hat{\beta} = 14.08$).	40
2.11	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.	42
2.12	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. .	43
2.13	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 $\hat{\pi}(\theta_{71} y_{71} = 4)$ for the rat tumor data. Panel (c) displays $\hat{\pi}(\theta_6 y_6 = 0)$ 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.	45
2.14	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 $\hat{\mathbb{E}}(\theta x)$ by the Gamma PEB, Robbins’ formula, Bayesian deconvolution, NPMLE, and our Elastic-Bayes estimate.	48
2.15	Demonstration of DS-Bayes with covariates on the Norberg insurance dataset. Panel (a) displays the U-function. Panel (b) shows the DS-prior (red), the PEB prior (blue) and the Kiefer-Wolfowitz NPMLE prior (green). Panel (c) shows macroinference with standard errors (using smooth bootstrap): two modes located at $0.57(\pm 0.094)$ and $1.41(\pm 0.261)$. Panels (d) through (f) show microinference for occupational groups 13, 22, and 53 (respectively).	50
2.16	The solid red line represents the DS-prior (without covariate) and the dashed green line is the DS-Prior when accounting for the covariate radius x_i	52

2.17	Comparisons of $DS(G, m)$ (red) with other empirical Bayes modeling cultures (green): (a) The DS-estimated prior is compared with the exponential prior model from Efron (1996); (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 $\hat{E}(\theta x)$ by Gamma PEB, Robbins' formula, Bayesian deconvolution, NPMLE, and our elastic-Bayes estimate.	55
2.18	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.	57
3.1	Modern data is data that is heterogeneous, large in volume, and low quality.	60
3.2	Tier 3 analysis for the physics and medical uncertainty data from Bailey (2017). Panel (a) and (c) shows the comparison density function overlaid on a histogram of the quantile values for the physics and medical uncertainty data, respectively. Panels (b) and (d) depict the original g (blue dashed), along with the maximum entropy (purple) and L2 (green) representation of the DS distribution.	66
3.3	For the Hubble data set, Panel (a) shows the uncertainty function which indicates structure change for the right tail, while (b) details $\hat{\theta}$ with a single mode at 63.7 ± 1.52 . Panel (c) shows the Tier 2 analysis, with DS-elastic Bayes shrinking toward the mode.	71
3.4	Tier 3 analysis for the Hubble constant data set. Panel (a) shows the comparison density. Panel (b) shows the histogram of the observed data along with the $\hat{f}(z; Z)$. Panel (c) displays the quantile as a function of $F(z)$	72
3.5	Tier 4 analysis of the Hubble constant data set. Panel (a) shows the posterior distribution of $\hat{\pi}(\theta y^*)$ where the posterior mean is 70.18 and the posterior mode is 70.11. Panel (b) is the Finite Bayes analysis, with mean of 70.22 and mode of 69.95. The 90% credible interval is (66.76, 73.62).	73
3.6	Third tier analysis of Big G's pairwise Z_{ij} with g as the t-distribution with 2 degrees of freedom and $m = 10$. Panel (a) shows the comparison densities for all pairs using $k = 48$ and $k = 47$. Panel (b) compares the DS distribution of $k = 48$ and $k = 47$, where the later shows only two significant modes. Panel (c) shows the difference in the quantile values for each distribution.	75

3.7	Panel (a) shows the pairwise comparisons of each study, where the colors reflect the groups identified through application of k-means clustering with the positive modes in Figure 3.6(b). The blue dashed line represents the first point where a measurement contains a value associated with the largest mode. Panel (b) shows a matrix plot of the groupings. The green portion of the plot represents the pairwise differences that are most extreme and coincide with the 1995-1996 measurements.	79
3.8	Tier 4 analysis of the Big G data set. Panel (a) shows the posterior distribution for $y^* = 6.67408$. The posterior mean and mode are both 6.6741. Panel (b) shows the finite Bayes inference, with mean of 6.6742 and mode of 6.6741. The 90% credible interval is (6.6738, 6.6744)	80
3.9	Tier 1 through 3 analysis of the ${}^7\text{Li}$ data. For Tier 1, $\hat{\pi} \equiv g$ thus Panel (a) shows one significant mode at 2.21 ± 0.009 . Panel (b) provides the Tier 2 analysis, which is an application of Stein's shrinkage. Panel (c) shows $f(z; Z)$ using comparisons to the median value. As with Tier 1, the Z_{ij} require no correction and thus $f(z; Z) \equiv t(2)$	81
4.1	An illustration of an extreme example in massively heterogeneous distributed data. Panel (a) shows the complete set of 5000 observations, while panel (b) gives an example of the 50 observations in a specific partition k . Panel (c) illustrates how simple averaging, fixed effect modeling, and the median of modes estimates the relationship based on partition estimates.	89
4.2	Panel (a) shows the U-function for the 'V-Data' problem, which tells us that there are two distinct modes. Panel (b) provides a comparison of the DS(G,m) prior versus the parametric prior for β . Panel (c) shows how the modes of (b) estimate the true nature of the data.	90
4.3	Scatter plots, regression lines, and $\hat{\beta}$ for three of the $k = 88$ stores in the cheese data. Each panel shows a different relationship between the promotional activity (percent display) and the sales volume.	92
4.4	The U-functions for the parameters of the cheese data, when analyzed under a distributed structure. Panel (a) is the intercept, (b) is the log(price), and (c) is the display.	94
4.5	A comparison of the distributions of the regression parameters between pooled data using Bayesian regression (panels (a) through (c)) and distributed data using DS prior (panels (d) through (e)). The comparison of panel (c) and panel (f) indicates the presence of Simpsons' paradox in the pooled analysis.	96
4.6	Three different candidates for $g(\theta)$: (a) no truncation; (b) zero truncation; and (c) double truncation.	104
4.7	Panel (a) shows estimates for the number of butterfly species caught in the following year $\hat{\mathbb{E}}(\theta y)$ by the Exponential PEB, Robbins' formula, Bayesian deconvolution, NPMLE, and our Elastic-Bayes estimate using the Exponential starting priors. Panel (b) shows the prediction of new number of species for additional years.	106

4.8 Prediction plots that compare results of DS Prior (g -modeling) with DS Distribution (f -modeling) for (a) butterfly and (b) Shakespeare's total lexicon. Panel (c) demonstrates the effectiveness of DS Distribution in using various percentages of sample size to predict the total words in Shakespeare's Hamlet. 114

CHAPTER 1

INTRODUCTION AND FOUNDATIONS

1.1 Introduction

Bayesians and frequentists have long been ambivalent toward one another, where the concept of “prior” represents the epicenter of this 250 years old tug-of-war (Efron, 1986; Sims, 2010; Stigler, 1982). Frequentists view the prior as a *weakness* that hampers scientific objectivity and corrupts the final statistical inference. Bayesians, though, view the prior as a needed *strength* that incorporates relevant domain-knowledge into the data analysis. While this conflict centered around the prior continues to rage, a bipartisan perspective seeks a unified approach that capitalizes on the advantages of both philosophies. With this perspective, the question naturally arises: how do we construct a consolidated Bayes-frequentist data analysis workflow that enjoys the best of both worlds? (Robbins, 1956; Good, 1992; Rubin, 1984; Efron, 2003) The objective of this dissertation is to provide the theoretical foundations and applications for one such modeling framework.

We begin by introducing the basic structure of our data model where we observe

samples $y = (y_1, \dots, y_k)$ from a known probability distribution $f(y | \theta)$, and the unobserved parameters $\theta = (\theta_1, \dots, \theta_k)$ are independent realizations from unknown $\pi(\theta)$:

$$y_i | \theta_i \stackrel{\text{ind}}{\sim} f(y_i | \theta_i), \quad i = 1, \dots, k$$
$$\Theta \sim \pi(\theta).$$

Given such a model, Bayesian inference typically aims at answering the following two questions:

- How should we combine k model parameters to come up with an overall, macro-level aggregated statistical behavior of $\theta_1, \dots, \theta_k$?
- Given the observables y_i , how should we simultaneously estimate individual micro-level parameters θ_i ?

These label these two inferential questions as ‘macroinference’ and ‘microinference.’ It is important to both clarify these terms and provide some examples of current practices.

The term macroinference refers to combining several, but possibly heterogeneous, estimates for a global inference. The primary objective of macroinference is to generate an estimate that more accurately and precisely describes the data than simply using a single estimate from a specific study. Pooled averages, Bayesian estimates, and the weighted mean all represent a sample of current methods that align with macroinference. The pooled average, although simple and straightforward, assumes an unlikely scenario where all of the observations are homogeneous thus ignores all heterogeneity and assumes there are no partition-specific $\{\theta_i\}$. In contrast, when we utilize Bayesian inferential thinking to arrive at consensus estimates, we incorporate heterogeneity into our estimate with a prior distribution that provides a functional structure for the unknown model parameters θ (Gelman et al., 2013). The weighted

mean also attempts to leverage the heterogeneity in the data through some form of weighting applied to the observations prior to combining for a grand estimate. John Tukey’s paper “Approximate Weight” (Tukey, 1948) explored such weighting, particularly the use of inappropriate weights when combining different sample means. Kafadar et al. (2003) points out that Tukey utilized the inverse variance as the optimal weight when determining a bound for the level of ‘inappropriateness’ in the weighting scheme. We see further evidence of the inverse variance as optimal weights in meta-analysis, which relies on such weighting to yield more precise estimates of θ (Hedges, 1983; Hedges and Olkin, 1985; Marin-Martinez and Sanchez-Meca, 2010). With the exception of the pooled average, all of these estimates require some assumed structure to account for any differences and heterogeneity in the observations. None of these methods, though, can provide a diagnostic as to the ‘correctness’ of their assumed structure. Therefore, there is limited justification as to which estimate provides the most accurate description of the observations.

‘MicroInference’ is a means to improve ‘local’ estimates for $\{\theta_i\}$. Specifically, the i^{th} value in a set of observations will have its own estimate θ_i given the observed data. Since the observed data used to find θ_i are based on the same or similar variables, we can improve the specific local estimate for θ_i by incorporating the information from the remaining observations. In Bayesian inference, microinference is also known as *shrinkage estimates*. Given $f(\theta | y)$ is the posterior distribution and λ_i lies between 0 and 1, we can define the *Bayes estimator* as:

$$\hat{\theta}^{(\text{Bayes})} = \mathbb{E}[f(\theta | y)] = \lambda_i \tilde{\theta}_i + (1 - \lambda_i) \mathbb{E}[\Theta] \tag{1.1.1}$$

Also known as Stein’s shrinkage formula (Stein, 1955), (1.1.1) ‘shrinks’ the partition proportions $\tilde{\theta}_i$ to the overall mean of the prior distribution $\mathbb{E}[\Theta]$. Because the parameters of the prior distribution are determined by the observed data, the ‘shrink-

age estimator' for a specific $\{y_i\}$ “borrows strength” (a term coined by John Tukey (Tukey, 1972; Brillinger, 2002; Efron, 2012)) from the rest of the observations while contributing its own information to the estimate.

The James-Stein estimator is one of the earliest examples of shrinkage estimators. In James and Stein (1961), the authors proved that, in some cases, the James-Stein estimator has a better quadratic loss than that of the MLE, meaning:

$$MSE_{\text{Stein}}(\hat{\theta}) < MSE_{\text{MLE}}(\hat{\theta}). \quad (1.1.2)$$

Further studies of shrinkage estimators can be found in Efron and Morris (1973), Efron and Morris (1977), Morris (1983), and Xie et al. (2016). Efron (2012) provides a wonderful and detailed history of shrinkage estimators, as well as how to utilize them in the modern era of large data. The shrinkage estimators represent the foundation of microinference and are a critical component in answering the second question.

In both macro- and micro-inference, we have expressed concerns about how current methods fail to diagnose the assumed structure of heterogeneity. We return to the fundamental model and focus on the distribution of θ : $\pi(\theta)$. As the prior distribution, $\pi(\theta)$ represents the functional structure of the heterogeneity of the observed data. To achieve the best possible estimates from macro- and microinferences, we must select an appropriate distribution for $\pi(\theta)$. The next section will introduce two primary philosophies on the prior distribution: traditional priors and empirical priors.

1.2 Priors

Because $\theta_i, i = 1, \dots, k$ are related through the prior distribution, it is a key part of Bayesian inference. The prior distribution allows θ to differ for each observed y , but also governs the variation and heterogeneity of θ . Gaining accurate macro- and microinference is contingent on selecting an appropriate prior $g(\theta)$ to estimate or

approximate the true prior distribution $\pi(\theta)$. The selection and identification of a ‘good’ prior is a topic of intense debate in Bayesian literature. Philosophies about the prior distribution lie between two extremes: the traditional approach and the empirical approach.

1.2.1 Traditional Approaches

Traditional approaches to determining a prior distribution requires an assumption on the functional form of g , but allows flexibility in the determination of its parameters.

Informative Prior Distributions

Since the prior distribution g represents the population distribution of potential parameter values θ , an informative prior is one that assumes a functional form of g that entails both our knowledge and uncertainty about θ (Gelman et al., 2013). In comparison, a *noninformative* prior is a diffuse prior that does not have a significant role in the posterior distribution (Gelman et al., 2013). In terms of our work, we ultimately want to capture a functional form for the heterogeneity and believe that informative priors provide the opportunity for a subject matter expert to encapsulate their knowledge. Furthermore, informative priors allow us to leverage conjugacy. From Gelman et al. (2013), if \mathcal{F} is a class of known probability distributions $f(y | \theta)$ and \mathcal{P} is a class of prior distributions $g(\theta)$, then \mathcal{P} is conjugate for \mathcal{F} if

$$p(\theta | y) \in \mathcal{P} \text{ for all } p(\cdot | \theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}. \quad (1.2.1)$$

Conjugate prior distributions are convenient because we know the distributional form of the posterior $f(y | \theta)$ will be the exact same as the prior $g(\theta)$.

Dirichlet-Process-based Approaches

A nonparametric Bayesian model defines the probability distribution of the prior on an infinite-dimensional space. The majority of work on Bayesian nonparametrics, particularly for binomial data problems, utilize a Dirichlet process prior on the space of all probability distributions (Liu, 1996). Let the characteristic measure of the Dirichlet process α be a finite measure on the interval $[0, 1]$. Then, a random probability measure P on that interval follows the Dirichlet process $\mathcal{D}(\alpha)$ if for every finite partition $\{A_1, \dots, A_m\}$ of $[0, 1]$, the random vector $\{P(A_1), \dots, P(A_m)\}$ follows a Dirichlet distribution with parameter $\{\alpha(A_1), \dots, \alpha(A_m)\}$.

The Dirichlet process priors (Ferguson, 1973a) allow the number of parameters of the model to grow along with the size of the data, which introduces a large amount of flexibility in the model. While a purely nonparametric method, it has some serious practical problems: it is computationally very expensive and always yields a discrete measure.

Weakly Informative Priors

A weakly informative prior is a proper prior, meaning that it does not depend on data and integrates to 1 (Gelman et al., 2013) that intentionally provides the least information than any available prior knowledge. As discussed in Gelman et al. (2008), these priors are a compromise between “fully informative prior distributions using application-specific information” and “noninformative priors, typically motivated by invariance principles.” Through a generic prior constraint which avoids specific information (Gelman et al., 2008), the weakly informative prior gives some structure on the uncertainty of the parameters while allowing the observed data more influence on inferences. In cases where the observed data provides evidence much stronger than that of a weakly informative prior, the data will dominate the inference (O’Hagan et al., 2006). Weakly informative priors enable approximations to more detailed

Bayesian analysis dependent on complex prior distributions (O’Hagan et al., 2006). Therefore, these priors are more appropriate for broader use than those more specific to a particular analysis (Gelman et al., 2008).

1.2.2 Empirical Approaches

The empirical approach is more formally known as ‘empirical Bayes.’ In empirical Bayes, the essential task is to learn the prior distribution from current statistical experience in lieu of relying on the assumptions prevalent in the traditional approach (Efron, 2013). It utilizes a combination of Bayesian and frequentist modeling strategies to estimate a prior’s unknown hyperparameters and/or structure using the observed data (Robbins, 1956; Efron, 2013). To avoid any complications or controversy surrounding the use of observed data for prior estimation *and* posterior inference, the prior is estimated by leaving out the observation targeted for posterior inference. Once estimated, the prior is then used in subsequent posterior analysis.

While parametric empirical Bayes establishes a parametric distribution for the prior and fixes the hyperparameters based on the data, nonparametric empirical Bayes makes no assumptions on the prior’s form and develops it based solely on the data (Robbins, 1956; Morris, 1983; Efron, 2016).

Parametric

From Morris (1983), parametric empirical Bayes requires the identification of a parametric distribution g for the prior, where the parameters for g are determined based on the data. Other methods (e.g. Efron (1996)) include techniques utilizing a class of priors that belong to the exponential family.

Nonparametric

Robbins (1956) introduced what many call the purest form of nonparametric empirical Bayes, which used an empirical distribution function G_{n-1} based on the observed data y to approximate the prior distribution g . Given observations $Y_i \sim f(y_i | \theta_i)$, we can approximate the unobserved values of θ_i using the observed values of y_i , and with the previous estimates of θ_i , get the empirical distribution function for θ :

$$G_{n-1}(\theta) = \frac{\text{number of } \theta_i \leq \theta}{n-1} \quad (1.2.2)$$

Robbins showed that, since $G_{n-1} \rightarrow G$ with probability 1 as $n \rightarrow \infty$, the Bayes estimate based on G_{n-1} , ψ_n , tends to the Bayes estimator $\nu_G(Y)$ of θ associated with the unknown prior distribution $G(\theta)$.

Recent works on nonparametric empirical Bayes include nonparametric maximum likelihood estimators (NPMLE) for mixture models (Koenker and Mizera, 2014; Koenker and Gu, 2017) and Bayesian deconvolution (Efron, 2016). NPMLE assumes the prior g is composed of a known density ψ , unknown parameters θ and an unknown mixture distribution $F(\theta)$. Kiefer and Wolfowitz (1956a) proposed using convex optimization to solve for $F(\theta)$, but solving the problem was computationally intensive with an extremely slow convergence in the EM algorithm (Koenker and Gu, 2017). In Koenker and Mizera (2014), the authors proposed a more efficient implementation using interior point methods and the optimization software MOSEK (Koenker and Mizera, 2014; Koenker and Gu, 2017). From NPMLE, the resulting distribution for g identifies significant mass points of the prior instead of a smooth distribution.

The Bayesian deconvolution method is purely nonparametric in that it makes no assumptions on the form of the prior distribution; instead, it allows the observed data to determine all aspects of g (Efron, 2016; Efron and Hastie, 2016). Efron's approach assumes that g consists of an exponential family with p parameters and a matrix

natural spline models with varying degrees of freedom (Efron, 2016, 2017). It uses a penalized maximum likelihood to estimate the unknown parameters of g , which then produces a smooth model of the prior distribution.

1.2.3 Which Philosophy is Best?

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 (Dempster, 1975; Berger, 1994) 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.

1.3 A Third Culture

Each EB modeling culture has its own strengths and shortcomings. For example, PEB methods are extremely efficient when the true prior is the conjugate prior g . On the other hand, NEB methods possess extraordinary robustness in the face of a miss-specified prior yet are inefficient when in fact $\pi \equiv g$. Acknowledge this trade-off,

Robbins (1980) proposed the following intriguing question: *how can this efficiency-robustness dilemma be resolved in a logical manner?* Essentially, Robbins question identifies an issue with current EB modeling techniques: while each framework performs optimal under specific conditions, neither are optimal when the conditions are unknown. To address this issue, the modeling framework must include a data analysis protocol that offers a mechanism to answer the following *intermediate* modeling questions prior to estimating $\hat{\pi}$:

1. Can we assess whether or not a selected prior g is adequate in light of the sample information?
2. In the event of a prior-data conflict, how do we estimate the correct functional form in a completely data-driven manner?

These questions encompass a formulation that relies on both Bayesian modeling principles and frequentist goodness-of-fit assessments. A unification of these two principles results in a third culture of empirical Bayes: generalized empirical Bayes (gEB). This third culture unites Bayesian and frequentist principles in addition to parametric and non-parametric philosophies. It provides an analyst with a graphical diagnostic tool to assess the worthiness of the selected prior g given the observed data. Furthermore, gEB incorporates a prior distribution that has the robustness of NEB but the flexibility to reduce to the PEB answer when g is appropriate; thereby turning Robbins' vision into action.

The core motivation behind gEB is more than just another recipe for estimating the prior from data. Instead, we seek to understand the mysterious prior in a transparent and definitive way. Critical to this understanding is to explore and answer the following questions:

- How can we provide automatic protection from unqualified specifications of the 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.

1.4 Summary of Contributions

This dissertation illuminates a path toward this goal with a methodology that is readily usable for wide-range of applied problems. We believe that generalized empirical Bayes can become an integral part of applied Bayesian modeling. Our approach includes some practical advantages over others, including:

- i. computational simplicity that does not require extensive Markov chain Monte Carlo (MCMC), variational methods, or other complex computational methods;
- ii. an easily interpretable framework whose generality easily encompasses a wide variety of common models;
- iii. readily applicable to a wide range of problems and easily integrated with current Bayesian analysis.

The next chapter introduces the theoretical foundation for the key component of gEB: the $DS(G, m)$ prior distribution. In addition to establishing its theoretical foundation and illustrating its capabilities through numerous applications, the chapter will demonstrate how gEB provides a visual diagnostic for situations when the assumed parametric prior g is not structurally adequate for the observed data. Chapter 3 demonstrates how the gEB framework can attack the heterogeneity problem

persistent in reproducibility, generalizability, and prediction. The key element of the chapter is the detailing of a new, non-parametric approach to meta-analysis. Chapter 4 includes several generalizations and applications of gEB. In addition to covariate analysis and a multi-faceted approach to the missing species problem, we extend gEB to the analysis of massively heterogeneous distributed data and discuss several key observations on current approaches. Finally, Chapter 5 concludes with two areas of future work: moving toward a full Bayesian approach and potential applications in cybersecurity.

CHAPTER 2

A NEW CLASS OF PRIOR DISTRIBUTION MODELS

2.1 The Model

The different stages of generalized empirical Bayes modeling is shown in Figure 2.1. It 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. Generalized empirical Bayes is an approach that will yield answers to all five of the phases using only a *single* algorithm. The foundation and innovation of the algorithm centers around a new family of prior distribution, the DS prior.

2.1.1 New Family of Prior Densities

This section serves two purposes. First, it provides a universal class of prior density models. Then, it provides the Fourier non-parametric representation of the models

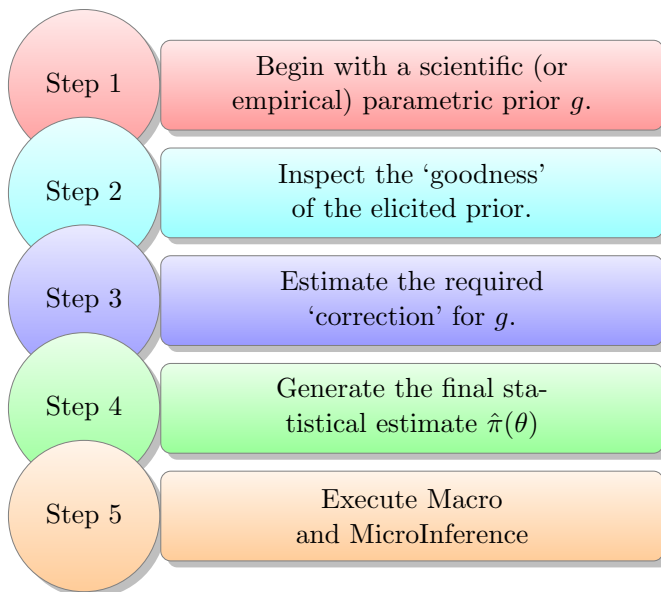


Figure 2.1: Proposed model-building approach for generalized empirical Bayes.

utilizing a specialized orthonormal basis.

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

$$\pi(\theta) = g(\theta; \alpha, \beta) d[G(\theta); G, \Pi], \quad (2.1.1)$$

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

The motivation behind the representation (2.1.1) 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 (2.1.1) always yields a proper density:

$$\int_{\theta} g(\theta) d[G(\theta); G, \Pi] d\theta = 1.$$

Also, this model specification has a unique *two-component* structure that combines the assumed parametric g with the d -function. The function d is a ‘‘correction’’ density to counter the possible misspecification bias of g . Ultimately, we can view the density function $d(u; G, \Pi)$ as a quantification for the ‘‘excess’’ *uncertainty* of the

assumed $g(\theta; \alpha, \beta)$. For that reason, we refer to $d(u; G, \Pi)$ as the *U-function*.

Since the square integrable $d[G(\theta); G, \Pi]$ lives in the Hilbert space $\mathcal{L}^2(G)$, we can approximate it by projecting into the orthonormal basis $\{T_j\}$ satisfying

$$\int T_i(\theta; G)T_j(\theta; G) dG = \delta_{ij}. \quad (2.1.2)$$

Here, δ_{ij} is zero if $i \neq j$ and 1 if $i = j$, which implies we can expand $d[G(\theta); G, \Pi]$ in a manner that is orthogonal to G . Now, the important question: how do we construct such a $T_j(\theta; G)$? To answer, we first need to introduce some notations and concepts.

Following Mukhopadhyay (2017), we define the concept of rank- G transform as the probability integral transformation of random variable X with respect to continuous measure G , as $G(X)$ where X is governed by the distribution F . Next, we construct $T_j(X; G)$, $j = 1, 2, \dots$, using the Gram Schmidt orthonormalization of the power of rank- G transformations of X , $\{G(X), G(X^2), \dots, G^j(X)\}$. As shown in Mukhopadhyay and Parzen (2014) and Mukhopadhyay (2017), the first five polynomial bases, which we will refer to as score functions, are:

$$T_0(X; G) = 1$$

$$T_1(X; G) = \sqrt{12}(G(x) - 0.5)$$

$$T_2(X; G) = \sqrt{5}(6G^2(x) - 6G(x) + 1)$$

$$T_3(X; G) = \sqrt{7}(20G^3(x) - 30G^2(x) + 12G(x) - 1)$$

$$T_4(X; G) = 3(70G^4(x) - 140G^3(x) + 90G^2(x) - 20G(x) + 1)$$

By design these polynomials are orthogonal with respect to the measure G , the distribution function of our selected prior distribution g . Note that $T_j(X : G)$ can be equivalently re-written as $\text{Leg}_j[G(X)]$, which are the shifted Legendre Polynomials

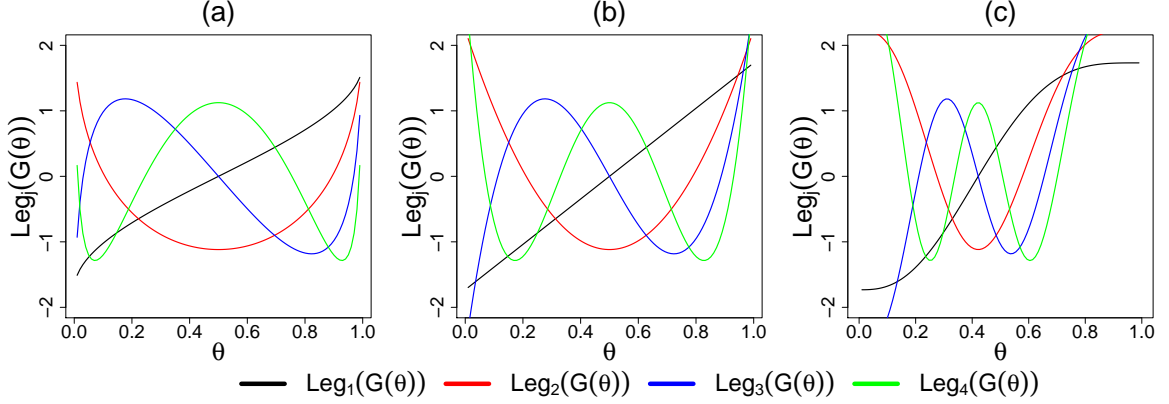


Figure 2.2: LP-polynomials $T_j(\theta; G_{\alpha,\beta})$ for family= "beta" with the following (α, β) choices: (a) Jeffrey's prior ($\alpha = \beta = 0.5$), (b) uniform prior ($\alpha = \beta = 1$), and (c) Beta($\alpha = 3, \beta = 4$). Note that for $U[0, 1]$ (the middle panel) $T_j \equiv \text{Leg}_j$, as $G(\theta)$ is simply θ in this case.

in $L^2[0, 1]$ evaluated at $G(X)$. Figure 2.2 shows $T_j = \text{Leg}_j(G_{\alpha,\beta}(\theta))$ for $j = 1, \dots, 4$ of three different parametric beta distributions.

This class of polynomials from the rank- G transform random variable are known as the ‘*LP family of rank polynomials*’ (Mukhopadhyay and Parzen, 2014; Mukhopadhyay, 2017). Note the ‘**L**’ plays a special role in nonparametric statistics that identifies robust methods based on rank and order statistics (e.g., Quantile-domain methods). The letter ‘**P**’ refers to *Polynomials*. Our T_j ’s are specifically-designed basis functions of the rank- G transform (*not* the raw x -values). Their unique structure enables the derivation of the estimator for d in the form of a linear combination of LP score functions of G . Therefore, we choose $T_j(\theta; G)$ to be $\text{Leg}_j \circ G(\theta)$, a member of the LP-class of rank-polynomials (Mukhopadhyay and Parzen, 2014).

The system $\{T_j\}$ possesses two attractive properties. First, they are polynomials of rank transform $G(\theta)$ and therefore constitute a robust basis. Second, they are orthonormal with respect to $\mathcal{L}^2(G)$ for *any* arbitrary G that is continuous. The score functions are not to be confused with standard Legendre polynomials $\text{Leg}_j(u)$, $0 < u < 1$, which are orthonormal with respect to Uniform $[0, 1]$ measure. The above discussion paves the way for the following definition.

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

$$\pi(\theta) = g(\theta; \alpha, \beta) \left[1 + \sum_{j=1}^m \text{LP}[j; G, \Pi] T_j(\theta; G) \right]. \quad (2.1.3)$$

where $T_j(\theta; G)$ is the j^{th} LP score function of G and

$$\text{LP}[j; G, \Pi] = \langle d, T_j \circ G^{-1} \rangle_{\mathcal{L}^2(0,1)} = \mathbb{E}[T_j(\Theta; G); \Pi]. \quad (2.1.4)$$

The LP-Fourier coefficients $\text{LP}[j; G, \Pi]$ are the key parameters that help us mathematically express the “gap” between a priori anticipated G and the true prior Π . When all the expansion coefficients are zero, we automatically recover g . This remarkable structure allows the $\text{DS}(G, m)$ to fully address Robbins’ concerns with regards to the compromise between the efficiency and robustness of empirical priors.

From Definition 2, when $\pi(\theta)$ is a member of $\text{DS}(G, m)$ class of priors then the orthogonal LP-transform coefficients (2.1.3) satisfy (2.1.4). Thus, given a random sample $\theta_1, \dots, \theta_k$ from $\pi(\theta)$, we could easily estimate the unknown LP-coefficients, and, thus, d and π , by computing the sample mean $k^{-1} \sum_{i=1}^k T_j(\theta_i; G)$. *Unfortunately, the θ_i ’s are unobserved.* Section 2.2 outlines an estimation strategy that appropriately and mathematically rectifies this dilemma. 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 2.1.2.

Under definition 2, we have $\text{DS}(G, m = 0) \equiv g(\theta; \alpha, \beta)$. The truncation point m in (2.1.3) reflects the *concentration* of permissible π around a known g . As $m \rightarrow \infty$, we include more basis functions that allows the $\text{DS}(G, m)$ prior to become more non-parametric (see Figure 2.3 for a pictorial description). 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

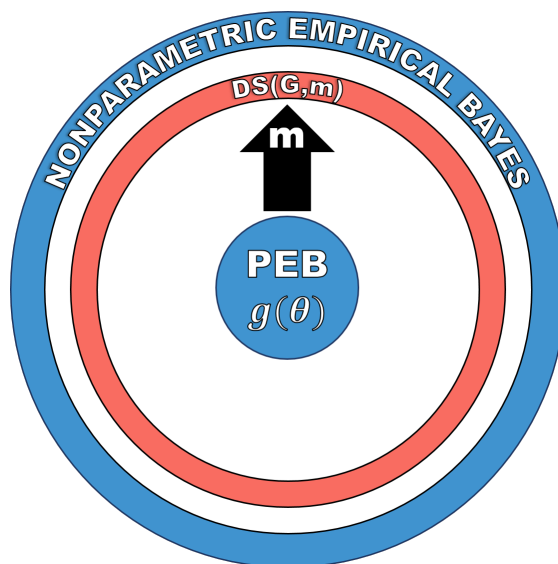


Figure 2.3: An illustration of how m allows the $DS(G, m)$ prior flexibility to deviate from $g(\theta)$, should the data indicate $g(\theta)$ requires correction.

around the domain-knowledge-based g by choosing m not “too big.” The first portion of Section 2.4 will demonstrate a quantitative approach to determining an optimal m based on established information criteria.

The motivations behind the name ‘DS-Prior’ are twofold. First, our formulation operationalizes I. J. Good’s ‘Successive Deepening’ idea (Good, 1983) for Bayesian data analysis:

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

Secondly, our prior has two components: A Scientific g that encodes an expert’s knowledge and a Data-driven d . That is to say that our framework embraces data and science, both, in a *testable* manner (Gelman et al., 2017).

2.1.2 Exploratory Diagnostics and U-Function

Is your data compatible with the pre-selected $g(\theta)$? If yes, we have our prior distribution and can avoid the arduous business of nonparametric estimation. If no,

we can model the “gap” between the parametric g and the true unknown prior π (often *far easier* than modeling π from scratch!). If the observed y_1, \dots, y_k look very unexpected given $g(\theta; \alpha, \beta)$, it is completely reasonable to question the self-selected prior. Here we provide a formal nonparametric exploratory procedure to comprehensively describe the uncertainty about the choice of g . Using the algorithm detailed in the next section, we estimate U-functions for four real data sets: the first three are binomial variate and the last one normal. The results are shown in Fig. 2.4.

- The rat tumor data set (Gelman et al., 2013) 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.
- The terbinafine data set (Young-Xu and Chan, 2008) 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 data set (Beckett and Diaconis, 1994) 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 set consists of forty randomized trials of a surgical treatment for stomach ulcers conducted between 1980 and 1989 (Sacks et al., 1990; Efron, 1996). Each of the 40 trials has an estimated log-odds ratio $y_i \mid \theta_i \sim \mathcal{N}(\theta_i, s_i^2)$ 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, MLE is not the only reasonable choice to estimate starting prior parameters. Other options may include

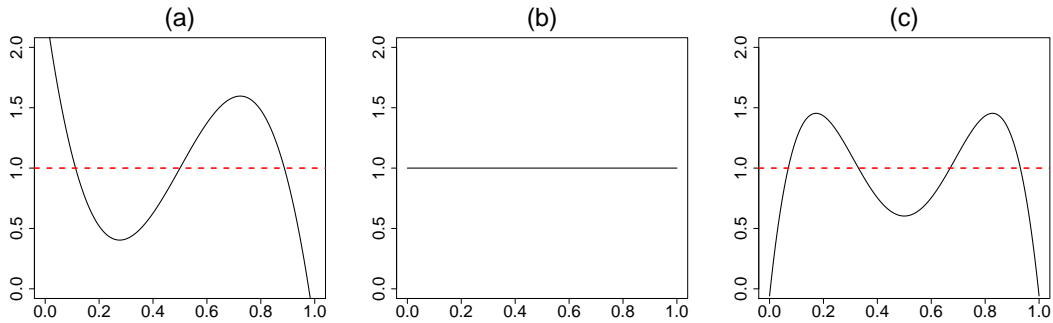


Figure 2.4: Graphical diagnostic tool: U-functions for (a) rat tumor data; (b) terbinafine and ulcer data; and (c) rolling tacks data. The deviation from uniformity (red dotted line) indicates that the default prior contradicts the observed data. The flat shape of the U-function in panel (b) suggests $\text{Beta}(1.24, 34.7)$ and $\mathcal{N}(-1.17, 0.98)$ are consistent with the terbinafine and ulcer data, respectively.

method of moments, other density modeling tools, and even subject-matter expert’s judgment. No matter the method one selects to get the initial parameter estimates for $g(\theta)$, the *shape* of the \hat{d} ; more specifically, its departure from uniformity, indicates the assumed conjugate prior $g(\theta; \alpha, \beta)$ needs ‘repair’ to resolve the prior-data conflict. For example, the flat shape of the estimated \hat{d} in Fig. 2.4(b) indicates that our initial selection of $g(\theta; \alpha, \beta)$ is appropriate for the terbinafine and ulcer data. Therefore, one can proceed with confidence using the parametric beta and normal prior respectively for conducting Bayesian inference.

In contrast, Figs. 2.4(a,c) provide a strong warning against 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, also known as the prior-data conflict, that exists between g and the observed data with 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 common Bayesian practice of assigning a hyperprior distribution on α and β is not always a fail-safe strategy and should be practiced with caution. The bottom line is uncertainty in the prior

probability model does not equate to uncertainty in parameters α and 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:

$$\text{qLP}(G \parallel \Pi) = \sum_j |\text{LP}[j; G, \Pi]|^2. \quad (2.1.5)$$

The motivation behind this definition comes from applying Parseval's identity, $\int_0^1 d^2(u; G, \Pi) = 1 + \text{qLP}(G \parallel \Pi)$, to (2.1.3). 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:*

$$\text{qLP}(G \parallel \Pi) \approx 2 \times \text{KL}(\Pi \parallel G), \quad (2.1.6)$$

where $\text{KL}(\Pi \parallel G)$ 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 $G(\theta) = u$:

$$\text{KL}(G \parallel \Pi) = \int \pi(\theta) \log \frac{\pi(\theta)}{g(\theta)} d\theta = \int_0^1 d(u; G, \Pi) \log d(u; G, \Pi) du. \quad (2.1.7)$$

Next, approximate $d \log d$ in (2.1.7) with the Taylor series expansion at

$$a = 1, (d - 1) + \frac{1}{2}(d - 1)^2:$$

$$\begin{aligned} \text{KL}(G \parallel \Pi) &= \int_0^1 (d(u; G, \Pi) - 1) + \frac{1}{2}(d(u; G, \Pi) - 1)^2 du \\ &= \int_0^1 d(u; G, \Pi) - 1 + \frac{1}{2}d(u; G, \Pi)^2 - d(u; G, \Pi) + \frac{1}{2} du \\ &= \frac{1}{2} \int_0^1 (d(u; G, \Pi)^2 - 1) du \\ &= \frac{1}{2} \int_0^1 d(u; G, \Pi)^2 du - \frac{1}{2} \end{aligned}$$

We know that applying Parseval's identity to (2.1.3) gives us

$$\int_0^1 d^2(u; G, \Pi) du = 1 + \text{qLP}(G \parallel \Pi), \text{ so it follows:}$$

$$\begin{aligned} 2 \times \text{KL}(G \parallel \Pi) &= 1 + \text{qLP}(G \parallel \Pi) - 1 \\ &= \text{qLP}(G \parallel \Pi) \end{aligned}$$

□

The U-Function provides a graphical measure of "goodness of fit" between the observed data and the selected prior that encourages "interactive" data analysis that is similar in spirit to Gelman et al. (1996). Subject-matter experts can use this tool to test and examine different hyperparameter choices in order to filter out the reasonable ones (see Chapter 4.2 for an illustration of this approach). This functionality might be especially valuable when multiple expert opinions are available. When \hat{d} shows evidence of the prior-data conflict, we need to provide information on what structure our starting parametric prior is missing. It is not enough to simply check the adequacy of g without informing the user an explanation for the misfit or what is the "deeper" structure absent in the starting parametric prior. Fortunately, our $\text{DS}(G, m)$ model suggests a simple, yet formal, guideline for upgrading:

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

where the shape of $\hat{d}(u; G, \Pi)$ captures the patterns which were not a priori anticipated. Hence our formalism *simultaneously* addresses the problem of uncertainty quantification and the subsequent model synthesis. The next Section outlines the process to estimate both the \hat{d} and $\hat{\pi}(\theta)$ without observing θ_i .

2.2 Estimation Method

2.2.1 Theory

In this Section, we lay out the key theoretical results that we use for designing our algorithm. In Chapter 1, we used the beta-binomial model as an example. Now, we want to take a more general approach in order to demonstrate the wide applicability of the DS prior. 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:

$$y_i | \theta_i \stackrel{\text{ind}}{\sim} f(y_i | \theta_i), \quad i = 1, \dots, k$$

$$\Theta \sim \text{DS}(G, m = 0).$$

Table 2.1 provides the marginal $f_G(y_i) = \int_{\theta_i} f(y_i | \theta_i)g(\theta_i) d\theta_i$ 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 $\mathbb{E}_G[h(\Theta_i) | y_i] = \int_{\theta_i} h(\theta_i)\pi_G(\theta_i | y_i) d\theta_i$. The subscript ‘ G ’ in these expressions underscores the fact that they are calculated for the conjugate g -model.

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

Table 2.1: 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 $[\pi_G(\theta_i y_i)]$
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($\alpha + y_i, \frac{\beta}{1+\beta}$)
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($\alpha + 1, \frac{\beta}{1+\beta y_i}$)

Theorem 2. Consider the following model:

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

where $\pi(\theta)$ is a member of DS(G, m) family (2.1.3), G being the associated conjugate prior. Under this framework, the following holds:

(a) The marginal distribution of y_i is given by

$$f_{\text{LP}}(y_i) = f_G(y_i) \left(1 + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G) | y_i] \right), \quad (2.2.1)$$

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

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

$$\pi_{\text{LP}}(\theta_i | y_i) = \frac{\pi_G(\theta_i | y_i) (1 + \sum_j \text{LP}[j; G, \Pi] T_j(\theta_i; G))}{1 + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G) | y_i]} \quad (2.2.2)$$

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

$$\mathbb{E}_{\text{LP}}[h(\Theta_i) | y_i] = \frac{\mathbb{E}_G[h(\Theta_i) | y_i] + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[h(\Theta_i)T_j(\Theta_i; G) | y_i]}{1 + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G) | y_i]} \quad (2.2.3)$$

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

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

Expanding the U-function in the LP-bases (2.1.3) yields

$$f_{\text{LP}}(y_i) = f_G(y_i) + \sum_j \text{LP}[j; G, \Pi] \int T_j(\theta_i; G) f(y_i | \theta_i) g(\theta_i; \alpha, \beta) d\theta_i. \quad (2.2.4)$$

The next step is to recognize that

$$f(y_i | \theta_i) g(\theta_i; \alpha, \beta) = f_G(y_i) \pi_G(\theta_i | y_i). \quad (2.2.5)$$

Substituting (2.2.5) in the second term of (2.2.4) leads to

$$\sum_j \text{LP}[j; G, \Pi] \int T_j(\theta_i; G) f(y_i | \theta_i) g(\theta_i; \alpha, \beta) d\theta_i = f_G(y_i) \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G) | y_i]. \quad (2.2.6)$$

Complete the proof of part (a) by replacing (2.2.6) into (2.2.4).

For part (b) of posterior distribution calculation we have

$$\pi_{\text{LP}}(\theta_i | y_i) = \frac{f(y_i | \theta_i) g(\theta_i; \alpha, \beta)}{f_{\text{LP}}(y_i)} \left\{ 1 + \sum_j \text{LP}[j; G, \Pi] T_j(\theta_i; G) \right\}. \quad (2.2.7)$$

Combine (2.2.1) and (2.2.5) to verify that

$$\frac{f(y_i | \theta_i) g(\theta_i; \alpha, \beta)}{f_{\text{LP}}(y_i)} = \frac{\pi_G(\theta_i | y_i)}{1 + \sum_j \text{LP}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G) | y_i]}. \quad (2.2.8)$$

Finish the proof of part (b) by replacing (2.2.8) into (2.2.7).

Part (c) is straightforward as

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

which is same as

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

by (2.2.2). Hence, result (2.2.3) is immediate. \square

Our LP-Bayes recipe (2.2.1)–(2.2.3) admits some interesting overall structure: the usual ‘parametric’ answer multiplied by a nonparametric correction factor involving $\text{LP}[j; G, \Pi]$ ’s. This decoupling pays dividends for theoretical interpretation as well as computation.

2.2.2 Algorithm

The critical parameters of our $\text{DS}(G, m)$ model are the LP-Fourier coefficients, which, as is evident from (2.1.4), we can simply estimate by their empirical counterpart $\widehat{\text{LP}}[j; G, \Pi] = k^{-1} \sum_{i=1}^k T_j(\theta_i; G)$. As we pointed out earlier, though, $\theta_1, \dots, \theta_k$ 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, \dots, y_k with distribution $f(y_i) = \int f(y_i | \theta_i) \pi(\theta_i) d\theta_i$. Following the spirit of the EM-algorithm, an obvious proxy for $T_j(\theta_i; G)$ would be its posterior mean

$\mathbb{E}_{\text{LP}}[T_j(\Theta_i; G) \mid y_i]$ that also naturally arises in the expression (2.2.1). This leads to the following ‘ghost’ LP-estimates:

$$\widetilde{\text{LP}}[j; G, \Pi] = k^{-1} \sum_{i=1}^k \mathbb{E}_{\text{LP}}[T_j(\Theta_i; G) \mid y_i], \quad (2.2.9)$$

satisfying $\mathbb{E}\{\widetilde{\text{LP}}[j; G, \Pi]\} = \widehat{\text{LP}}[j; G, \Pi]$ ($j = 1, \dots, m$), by virtue of the law of iterated expectations. We continue to refine these estimates until the difference between iterations is less than some ϵ . The following algorithm implements this strategy.

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

Step 0. Input: Data (y_1, \dots, 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: $\text{LP}^{(0)}[j; G, \Pi] = 0$ for $j = 1, \dots, m$. For iteration $\ell > 0$, perform steps (2-3) until convergence: $\sum_{j=1}^m |\widetilde{\text{LP}}^{(\ell)}[j; G, \Pi] - \widetilde{\text{LP}}^{(\ell-1)}[j; G, \Pi]|^2 \leq \epsilon$.

Step 2. Compute $\mathbb{E}_{\{\ell-1\}}[T_j(\Theta_i; G) \mid y_i]$ by substituting $\{\widetilde{\text{LP}}^{(\ell-1)}[j; G, \Pi]\}_{j=1}^m$ into (2.2.3), where $h(\theta_i) = \text{Leg}_j \circ G(\theta_i)$.

Step 3. Determine the ‘ghost’ LP-estimates:

$$\widetilde{\text{LP}}^{(\ell)}[j; G, \Pi] = k^{-1} \sum_{i=1}^k \mathbb{E}_{\{\ell-1\}}[T_j(\Theta_i; G) \mid y_i] \quad (j = 1, \dots, m).$$

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

For a more enhanced and parsimonious answer, we want the model to include only the statistically significant LP coefficients from the output. This “smoothing” process requires the use of Schwartz’s BIC-based model-selection criteria to deter-

mine significantly non-zero LP-coefficients. We arrange $\widehat{\text{LP}}[j; G, \Pi]$'s in a decreasing magnitude and choose m that maximizes

$$\text{BIC}(m) = \sum_{j=1}^m |\widehat{\text{LP}}[j; G, \Pi]|^2 - \frac{m \log(k)}{k}. \quad (2.2.10)$$

It is also possible to “smooth” with AIC-based model selection criteria. As with BIC, arrange $\widehat{\text{LP}}[j; G, \Pi]$'s in a decreasing magnitude and choose m that maximizes

$$\text{AIC}(m) = \sum_{j=1}^m |\widehat{\text{LP}}[j; G, \Pi]|^2 - \frac{2m}{k}. \quad (2.2.11)$$

Due to the less-restrictive nature of AIC as compared to BIC, the AIC-smoothed models tend to have more LP coefficients and may have more ‘extreme’ functional forms. As with selecting an appropriate m , selecting the smoothing criteria depends on the amount of flexibility the user wishes to include in his or her model.

2.2.3 Maximum Entropy Representation

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

$$\check{\pi}(\theta) = g(\theta; \alpha, \beta) \exp \left[c_0 + \sum_j c_j T_j(\theta; G) \right], \quad (2.2.12)$$

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 $\widehat{\text{LP}}[j; G, \Pi]$, $j =$

$1, \dots, m$.

Step 1. Define the set $\mathcal{J} = \{j : |\widehat{\text{LP}}[j; G, \Pi]| > 0\}$, 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 $\check{\pi}(\theta)$ of (2.2.12), solve the following sets of moment equality constraints:

$$\widehat{\text{LP}}[j; G, \Pi] = \int T_j(\theta; G) \check{\pi}(\theta) d\theta, \quad \text{for } j \in \mathcal{J}. \quad (2.2.13)$$

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

Two Data Examples. Here we carry out the maximum entropy analysis for the rat data (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 (De Blok et al., 2001).

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

$$\check{\pi}(\theta) = g(\theta; \alpha, \beta) \exp[-0.13 - 0.52T_3(\theta; G)]. \quad (2.2.14)$$

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

$$\check{\pi}(\theta) = g(\theta; \mu, \tau^2) \exp[-0.15 + 0.26T_3(\theta; G) - 0.28T_4(\theta; G) + 0.46T_5(\theta; G)]. \quad (2.2.15)$$

The resulting LP-maximum-entropy $\text{DS}(G, m)$ priors are shown in Figure 2.5. 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.

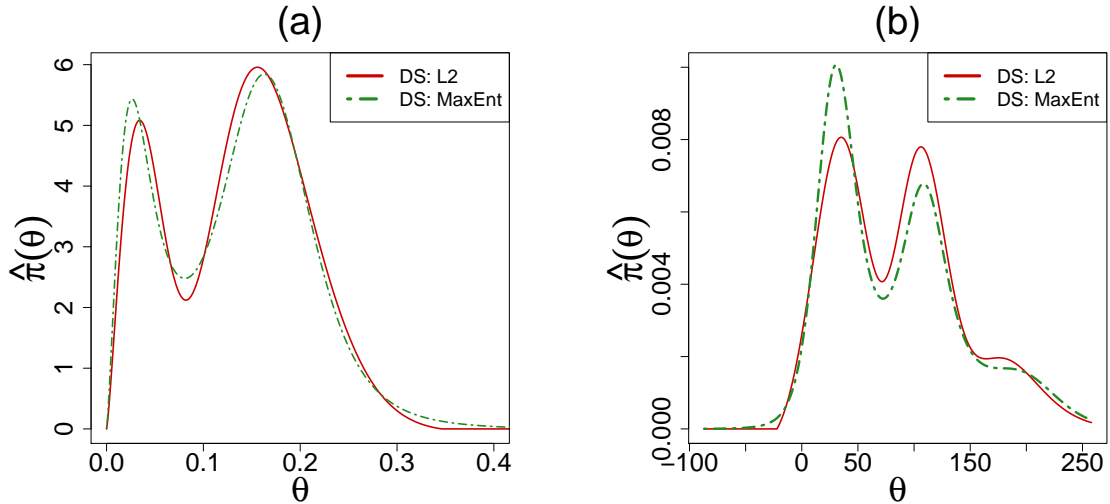


Figure 2.5: Comparison of \mathcal{L}^2 (solid red line) and maximum entropy (two-dash green line) estimates of DS prior. Panel (a) shows the comparison for the rat tumor data, while panel (b) illustrates the difference (in modal shapes) for the galaxy data.

2.2.4 Results

In addition to the rat tumor data (cf. Section 2.1.2), here we introduce and analyze three additional datasets: two binomial and one Poissonian example.

- The surgical node data involves number of malignant lymph nodes removed during intestinal surgery (Efron, 2016; Efron and Hastie, 2016). 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 consists of $k = 5$ samples of the number of defects y_i found in $n_i = 5$ lots of welding material (Martz and Lian, 1974).
- The insurance data, shown in Table 2.4, provides a single year of claims data for an automobile insurance company in Europe (Efron and Hastie, 2016). The counts $y_i \sim \text{Poisson}(\theta_i)$ represent the total number of people who had i claims in a single year.

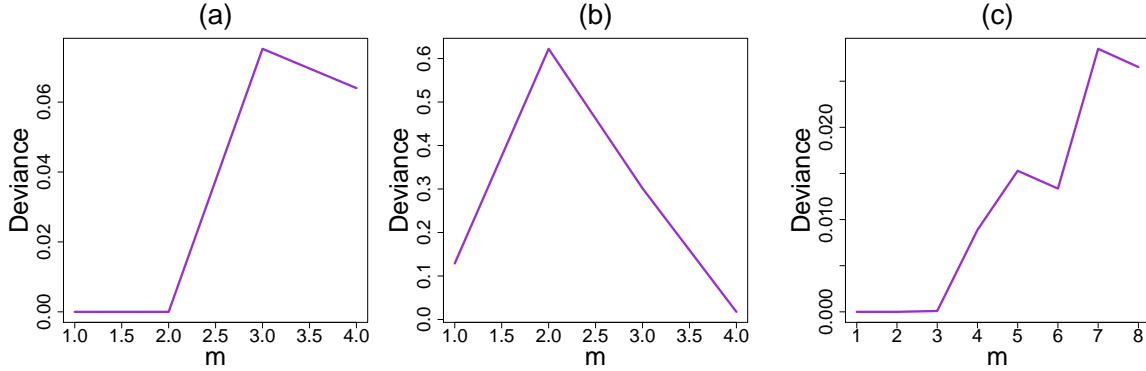


Figure 2.6: Plots of penalized uncertainty quantification with respect to m for (a) rat tumor data, (b) shipyard data, and (c) surgical node data. The peak bend in each plot indicates the optimal m for each data set.

We begin by demonstrating how to determine an appropriate m for three of these data sets. As mentioned in the previous section, we want to select m so that it provides enough flexibility for the U-function to adjust the parametric prior g . We use the penalized version of the uncertainty quantification shown in (2.2.10) for the rat tumor and Navy Shipyard data sets and (2.2.11) for the surgical node data set. Figure 2.6 provides the ‘deviance plots’ for the three data sets. In each of the plots, we see a peak-shape that identifies the value of m which maximizes the penalized deviances for the respective data sets. In most cases, selecting the m that corresponds to the maximum deviance is appropriate for the given data set. There may arise a time when a user would select an m that corresponds to a peak which is not necessarily a maximum, such as $m = 5$ for the surgical node data. These situations typically correspond to situations where the size of the data set is not appropriate to the maximum selected m .

Figure 2.7 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 $\alpha = 2.30$, $\beta = 14.08$:

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)[1 - 0.50T_3(\theta; G)]. \quad (2.2.16)$$

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

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)[1 - 0.07T_3(\theta; G) - 0.11T_4(\theta; G) + 0.09T_5(\theta; G) + 0.13T_7(\theta; G)]. \quad (2.2.17)$$

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

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)[1 - 0.67T_1(\theta; G) + 0.90T_2(\theta; G)]. \quad (2.2.18)$$

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

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)[1 - 0.26T_2(\theta; G)]. \quad (2.2.19)$$

The rat tumor data shows a prominent bimodal shape, which does not come as a surprise in light of Fig. 2.4(a). For the surgical data, the DS-prior puts excess mass around 0.4 which concurs with the findings of Section 4.2 in Efron (2016). In the case of the Navy shipyard 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 $\hat{\pi}$ (2.2.19), which makes it slightly wider in the middle with sharper peak and tail.

2.3 Inference

2.3.1 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*

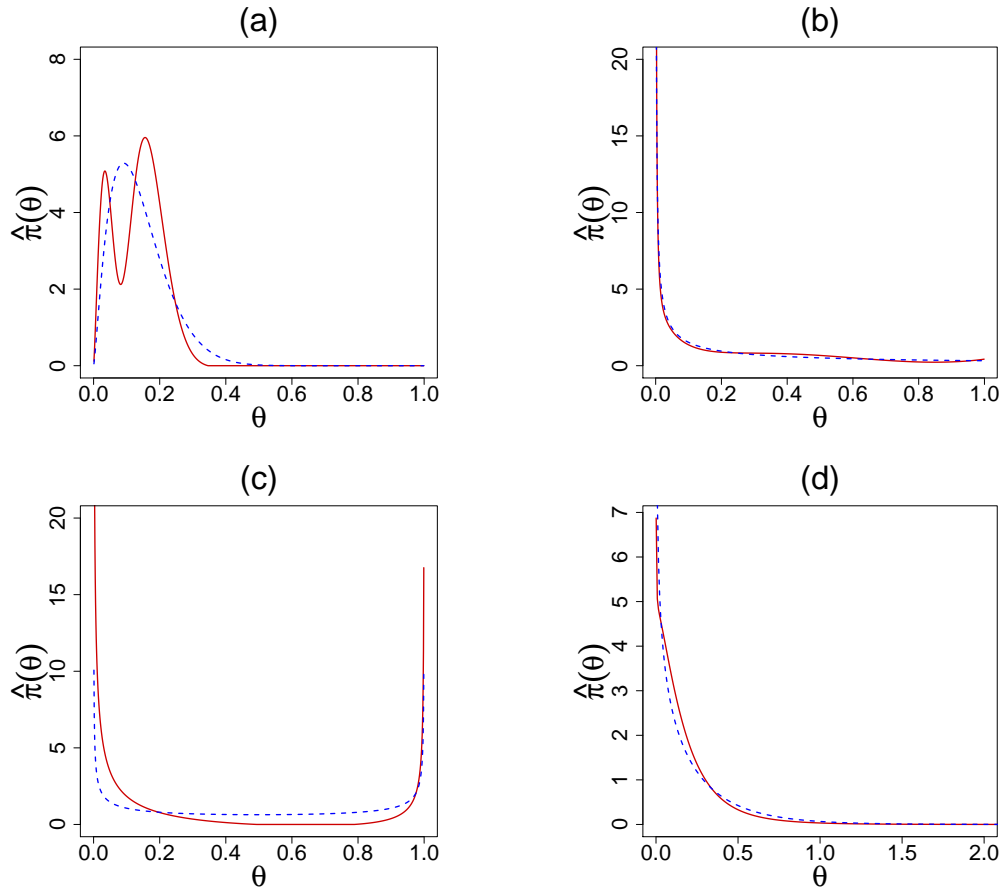


Figure 2.7: Comparisons of the $DS(G, m)$ prior $\hat{\pi}(\theta)$ (solid red) with the respective parametric EB (PEB) priors $g(\theta; \alpha, \beta)$ (dashed blue) for the (a) rat tumor data, (b) surgical node data, (c) Navy shipyard data, and (d) insurance data.

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. A critical component of the macro-level inference is the standard error (SE). We will utilize a smooth non-parametric bootstrap sampling method that will generate samples from the $DS(G, m)$ model through an accept/reject sampling scheme (Mukhopadhyay, 2017):

$DS(G, m)$ Sampling Algorithm

Step 1. Generate Θ from g ; independent of Θ , generate U from $\text{Uniform}[0, 1]$.

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

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

otherwise, discard Θ and return to Step 1.

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

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

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. 2.4(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 $\text{DS}(G, m)$ -sampler with $m = 0$ to compute the SE: 0.034 ± 0.006 , highlighted in the Fig. 2.8(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. The following examples illustrate other ways heterogeneity can cloud possible estimates.

Rat tumor and rolling tacks data analysis. Can we always extract a “single” overall number to aptly describe k parallel studies? Unfortunately, no. In order to appreciate this, let us look at Figs. 2.8 (a,c), which depict the estimated DS-prior for the rat tumor and rolling tacks data. We highlight two key observations:

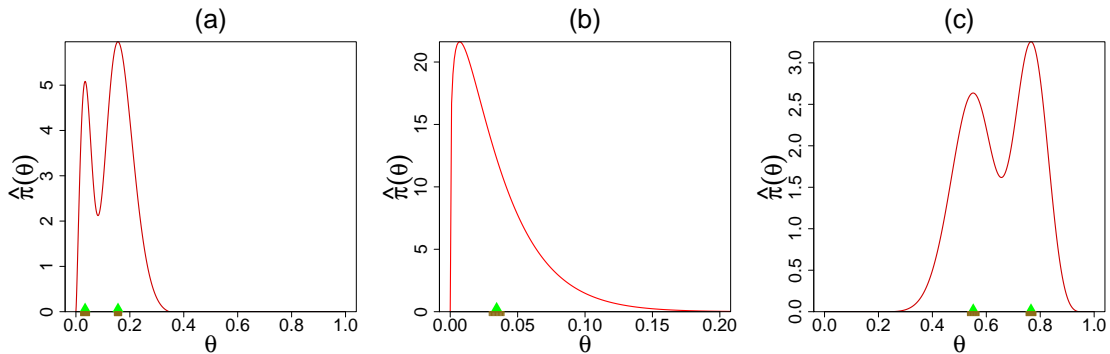


Figure 2.8: Estimated macro-inference summary along with standard errors (using smooth bootstrap) are shown. Panel (a) displays the rat tumor data modes located at $0.034 (\pm 0.016)$ and $0.156 (\pm 0.016)$. Panel (b) shows the estimated unimodal prior of the terbinafine data has a mean at $0.034 (\pm 0.006)$. Panel (c) presents the modes of the rolling tacks data at $0.55 (\pm 0.022)$ and $0.77 (\pm 0.018)$.

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 Liu (1996). 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 (Tarone, 1982; Dempster et al., 1983; Gelman et al., 2013). The resulting two groups of rat tumor experiments are enumerated in the Table 2.2. 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.

Table 2.2: Two group partitions of the rat tumor studies based on K-means clustering on the posterior mode predictions (see Section 2.3.3 and Fig. 2.10(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)

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 in overestimating one group and underestimating another. We prefer *modes* of $\hat{\pi}(\theta)$, 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 (Cox, 1990; Efron, 1996). 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, we recommend group-specific modes instead of the prior-mean summary. Our algorithm is also capable of finding data-driven clusters of the partially exchangeable studies in a fully automated manner.

2.3.2 Learning From Uncertain Data

The fields of metrology, chemistry, physics, biology, and engineering routinely work with a set of measurements made by k different laboratories in the form of y_1, \dots, y_k along with their estimated standard errors s_1, \dots, s_k . Given this uncertain data, a fundamental problem of interest is the final inference concerning (i) the estimation of the consensus value of the measurand, and (ii) the evaluation of its associated

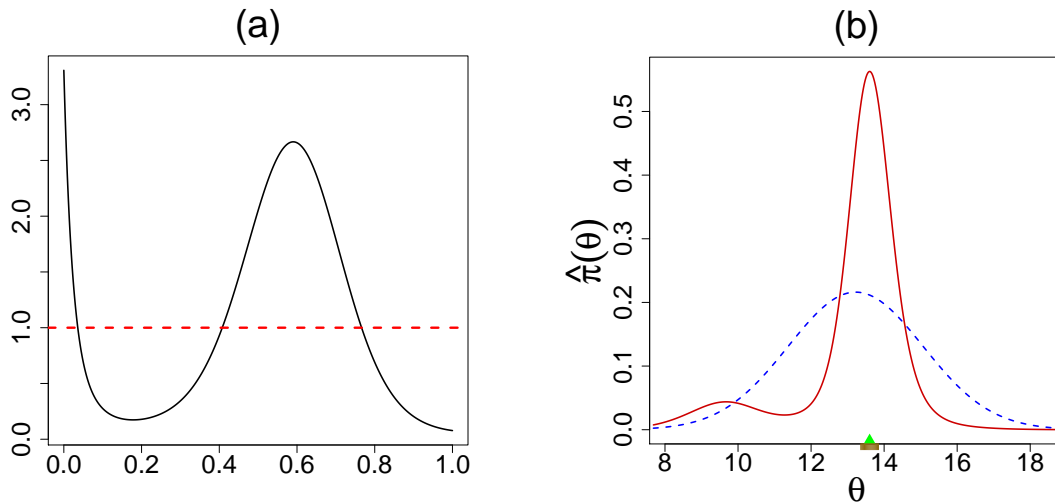


Figure 2.9: Panel (a) shows the U-function, while panel (b) compares the DS-prior $\hat{\pi}(\theta)$ (solid red) with the PEB prior $g(\theta; \alpha, \beta)$ (dashed blue) for the arsenic data. Based on the estimated macro-inference summary along with standard errors (using smooth bootstrap), the best consensus value is the mode 13.6 (± 0.242).

uncertainty. The data in Table 2.3 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 Administrations National Status and Trends Program Ninth Round Intercomparison Exercise (Willie and Berman, 1995).

Table 2.3: Measurements (sorted) along with their uncertainty from different laboratories in arsenic data.

Laboratory	1	2	3	4	5	...	25	26	27	28
Measurement (y_i)	9.78	10.18	10.35	11.60	12.01	...	14.70	15.00	15.10	15.50
Uncertainty (s_i)	0.30	0.46	0.07	0.78	2.62	...	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 $\Theta_i \sim \text{DS}(G, m)$ ($i = 1, \dots, 28$) with G being $\mathcal{N}(\mu, \tau^2)$. The shape of the estimated U-function in Fig. 2.9(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 (Rukhin and Vangel, 1998). The

DS-corrected $\hat{\pi}$ shows an interesting asymmetric pattern with two distinct modes. The left mode represents measurements from three laboratories that are unlike the majority. The result of our macro-inference is shown in Fig. 2.9(b), which delivers the consensus value 13.6 ± 0.24 . This value is clearly far more resistant to fairly extreme low measurements and surprisingly 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. In another analysis of the arsenic data, Possolo (2013) fits a Bayesian hierarchical model with prior as Students t_ν where the degrees of freedom were also treated as a random variable over some arbitrary range $\{3, \dots, 118\}$. 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. While key to measurement sciences, distinguishing long-tail behaviour from bimodality is an important problem of applied statistics by itself. Chapter 3 further explores this topic, detailing how gEB and the $DS(G, m)$ prior provide effective tools for uncertain data analysis.

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 (Toman and Possolo, 2009), and finally (v) our theory is still applicable for very small number of parallel cases (cf. Fig. 2.7(c)), a situation which is not uncommon in inter-laboratory studies.

2.3.3 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 are at least three ways to answer this question:

- Frequentist MLE estimate: An obvious estimate would be the sample proportion $\tilde{\theta}_i : y_{71}/n_{71} = 0.286$. 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 $\check{\theta}_i = \mathbb{E}_G[\Theta_i|y_i]$ operationalizes this heuristic, which in the Binomial case takes the following form:

$$\check{\theta}_i = \frac{n_i}{\alpha + \beta + n_i} \tilde{\theta}_i + \frac{\alpha + \beta}{\alpha + \beta + n_i} \mathbb{E}_G[\Theta]. \quad (2.3.1)$$

This is famously known as Stein’s shrinkage formula (Stein, 1955; Efron and Morris, 1975), 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 $\tilde{\theta}_i$ ’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. 2.8(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

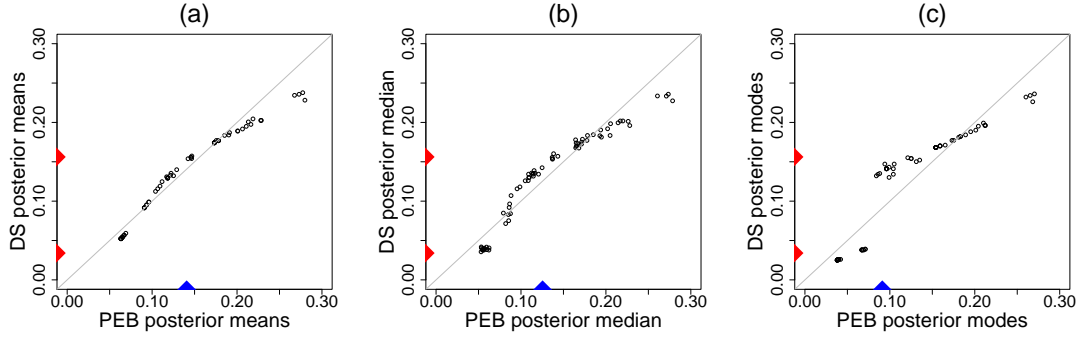


Figure 2.10: Comparisons of DS Elastic-Bayes and PEB posterior predictions of the rat tumor data: (a) posterior means, (b) posterior medians, and (c) posterior modes. The vertical red triangles indicate the location of the modes on the DS prior; the blue triangles respectively denote the mean, median, and mode of the parametric Beta($\hat{\alpha} = 2.3, \hat{\beta} = 14.08$).

full dataset–*selective shrinkage*, e.g., for the rat data, it would be the group 2 of Table 2.2. This brings us to the question: how can we rectify the parametric empirical Bayes estimate $\check{\theta}_i$ so that it accounts for structured heterogeneity? The formula (2.2.3) gives us the required (nonlinear) adjusting factor:

$$\hat{\theta}_i = \frac{\check{\theta}_i + \sum_j \widehat{\text{LP}}[j; G, \Pi] \mathbb{E}_G[\Theta_i T_j(\Theta_i; G)|y_i]}{1 + \sum_j \widehat{\text{LP}}[j; G, \Pi] \mathbb{E}_G[T_j(\Theta_i; G)|y_i]}, \quad (2.3.2)$$

Equation (2.3.2) dictates the magnitude and direction of shrinkage in a completely data-driven manner via the LP-Fourier coefficients. When $d \equiv 1$, i.e., all the $\text{LP}[j; G, \Pi]$ are zero, (2.3.2) reproduces the parametric $\check{\theta}_i$. 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 adjusts it based on the data.

Rat tumor example. Figure 2.10 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 $\tilde{\theta}_i$ 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. 2.10 (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 nonparametric 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. 2.11(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. 2.11 (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.

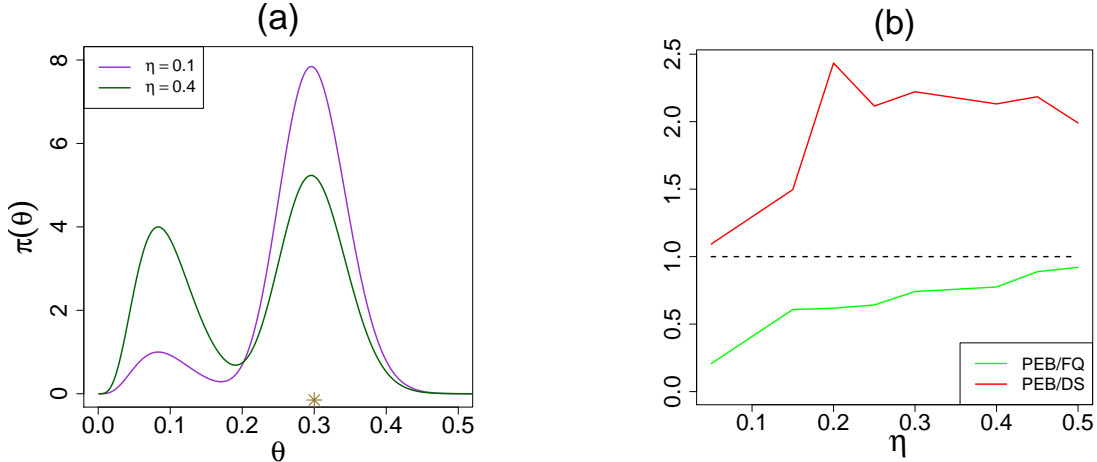


Figure 2.11: Panel (a) illustrates the prior-data conflict for $\eta = 0.1$ versus $\eta = 0.4$; ‘*’ denotes 0.3, the true mean of y_{new} . Panel (b) shows the MSE ratios for PEB to frequentist MLE (PEB/FQ; green) and PEB to DS (PEB/DS; red) with respect to η . Notice that as more prior-data conflict is introduced, DS outperforms PEB while frequentist MLE performance improves.

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 trials, where the heterogeneity arises due to changing conditions.

The Robbins’ Puzzle. Now, we will use two simulated scenarios to compare the $\text{DS}(G, m)$ prior with with two recent methods: Efron’s Bayesian deconvolution (implemented in the `deconvolveR` package), and Koenker’s NPMLE (implemented in the `REBayes` package).

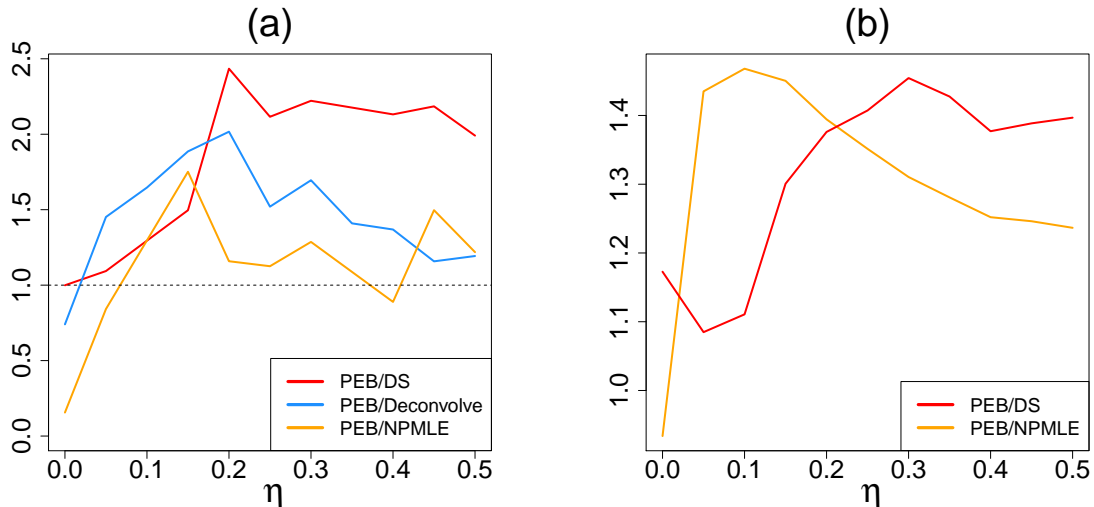


Figure 2.12: Results of two separate simulations comparing DS with other methods. In (a), the MSE ratios for PEB to empirical Bayes deconvolution (PEB/Dec; blue), PEB to Kiefer-Wolfowitz NPMLE using REBayes Bmix (PEB/NPMLE; orange) and PEB to DS (PEB/DS; red) with respect to η . Panel (b) shows the ratio of empirical risks after applying both DS and NPMLE methods to Robbins’ ‘compound decision’ problem.

Example 1. Here we operate under the exact settings presented in Pharma-example. Figure 2.12(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. As expected, the Stein’s parametric shrinkage estimator dominates other nonparametric approaches. On the other hand, $\eta = 0.5$ represents a complicated situation where $\pi(\theta) = \frac{1}{2}\text{Beta}(5, 45) + \frac{1}{2}\text{Beta}(30, 70)$. 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 . This result is another example how gEB elegantly addresses the robustness-efficiency puzzle of Robbins (1980).

Example 2. Next, we investigate the prediction problem originally introduced by Robbins (1951) and discussed in Gu and Koenker (2016). We observe $Y_i = \theta_i + \epsilon_i$, $i = 1 \cdots k$, where $\epsilon_i \stackrel{\text{ind}}{\sim} \mathcal{N}(0, 1)$, 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 $L(\hat{\theta}, \theta) = k^{-1} \sum_{i=1}^k |\hat{\theta}_i - \theta_i|$. 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 2.12(b) shows a very interesting result: Kiefer-Wolfowitz NPMLE method performs significantly better than the DS-elastic Bayes when $0 < \eta < .2$. 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.

Four additional real examples. Figure 2.13 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 2.13(a) shows posterior distributions for three selected patients, which are indistinguishable from Efron’s deconvolution answer (Cox and Efron, 2017, 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. 2.13(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.

[†]Mean loss is computed over 500 replications.

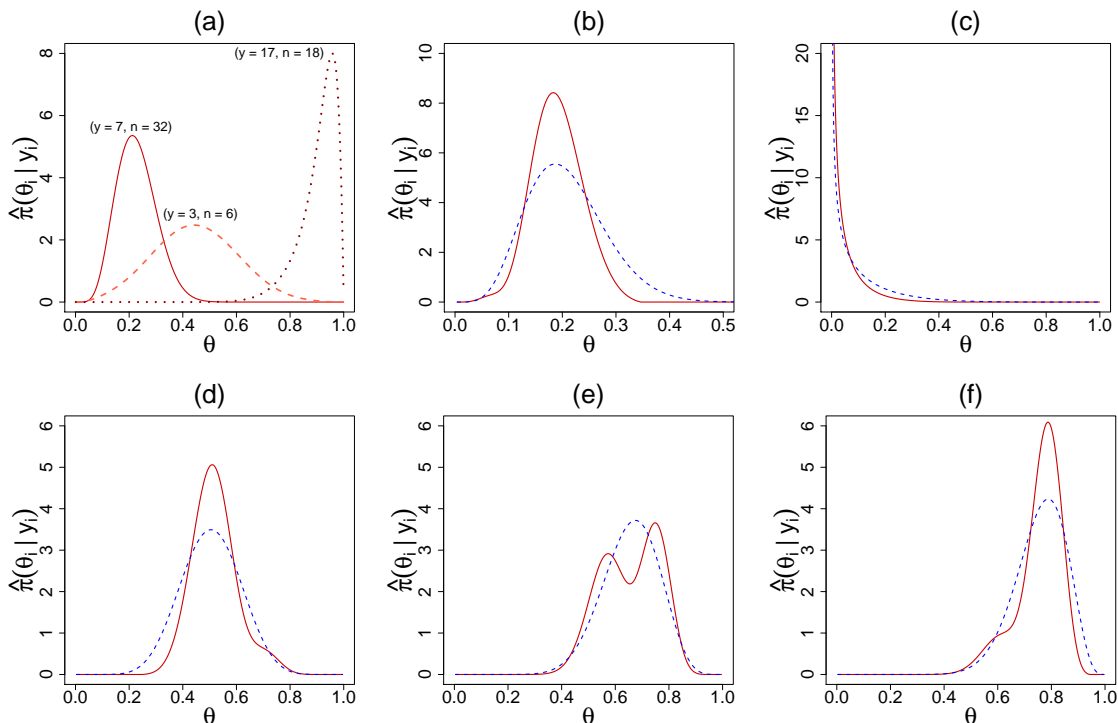


Figure 2.13: Panel (a) shows DS posterior plots of three observations from the surgical node data: $(y = 7, n = 32)$, $(y = 3, n = 6)$, and $(y = 17, n = 18)$. For panels (b) through (f), red denotes the DS posterior and blue dashed is the PEB posterior. Panel (b) is $\hat{\pi}(\theta_{71}|y_{71} = 4)$ for the rat tumor data. Panel (c) displays $\hat{\pi}(\theta_6|y_6 = 0)$ 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.

2.13 (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 (1993).

2.3.4 Poisson Smoothing: The Two Cultures

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

$$\mathbb{E}[\Theta|Y = y] = \frac{\int_0^\infty \theta [e^{-\theta} \theta^y / y!] \pi(\theta) d\theta}{\int_0^\infty [e^{-\theta} \theta^y / y!] \pi(\theta) d\theta}; \quad y = 0, 1, 2, \dots \quad (2.3.3)$$

There are two primary approaches for estimating (2.3.3):

- Parametric Culture (Fisher et al., 1943; Maritz, 1969): 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 (Robbins, 1956; Efron, 2003; Gu and Koenker, 2016): This was born out of Herbert Robbins’ ingenious observation that (2.3.3) 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 (Efron, 2003) and the Kiefer-Wolfowitz non-parametric MLE (NPMLE) (Kiefer and Wolfowitz, 1956b; Gu and Koenker, 2016). Efron’s technique can be viewed as *smooth* nonparametric approach, whereas NPMLE generates a discrete (atomic) probability measure.

Table 2.4: 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 2.4 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 DS elastic-Bayes. 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 2.2.4) the PEB estimate needs a second-order correction to resolve the discrepancy between the Gamma prior and data. The LP-nonparametric Stein estimates are shown in the last row of Table 2.4.

The butterfly data. The next example is Corbet’s Butterfly data (Fisher et al., 1943)– 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 $y = 1, \dots, 24$. Figure 2.14(b) plots different Bayes estimates. The Robbins procedure suffers from ‘roughness’ similar to its performance in the insurance data in Table 2.4. The blue dotted line represents the linear PEB estimate with $\alpha = 0.104$ and $\beta = 89.79$ (same as Eq. 6.24 of Efron and Hastie (2016)) 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 (Robbins, 1980). We will further explore the application of generalized empirical Bayes and the $DS(g, m)$ model to the missing species problem in Chapter 4.2.

2.4 Incorporating Covariates

We present two examples detailing how the generalized empirical Bayes framework can easily accommodate additional covariates. First, we will use a new insurance

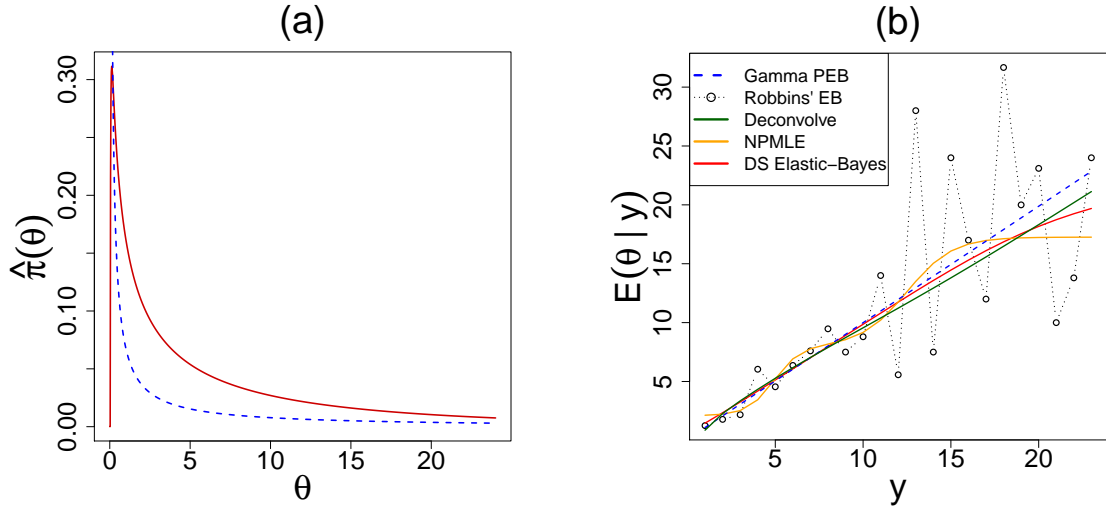


Figure 2.14: Panel (a) displays the estimated DS($G, m = 4$) prior (solid red) with the PEB Gamma prior $g(\theta; \alpha, \beta)$ (dashed blue) for the butterfly data; these results indicate that Fisher’s Gamma-prior guess required some correction. Panel (b) shows estimates for the number of butterfly species caught in the following year $\hat{E}(\theta | x)$ by the Gamma PEB, Robbins’ formula, Bayesian deconvolution, NPMLE, and our Elastic-Bayes estimate.

claims data set to illustrate how one incorporates an exposure into a Poisson-Gamma model. The second example is a Normal-Normal model for the rotational velocities of specific galaxies. In this example, we compare results using only the rotational velocity data to results that include the galaxy’s radius as a covariate.

2.4.1 The Norberg Example

. The Norberg insurance dataset (Norberg, 1989) 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 (1989), we assume $Y_i \sim \text{Poisson}(\theta_i E_i)$. Given the normalized E_i , we interpret θ_i as the occupational-specific rate of risk per contract period.

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

$$\hat{\pi}(\theta) = g(\theta; \alpha, \beta)[1 - 0.70T_1(\theta; G) + 0.83T_2(\theta; G) - 0.53T_3(\theta; G)]. \quad (2.4.1)$$

In Figure 2.15(a), the U-function clearly indicates potential prior-data conflict when using $\pi(\theta) = \text{Gamma}(6.02, 0.20)$. Figure 2.15(b) displays the DS prior (red) along with the parametric EB (blue) and the Kiefer-Wolfowitz NPML 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 2.15(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 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 2.15(d), we have an occupational category that identifies as higher risk with a small lower risk component. The unimodality in Figure 2.15(e) clearly indicates that category is a higher risk of claim based on exposure. Finally, the occupational category in Figure 2.15(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.

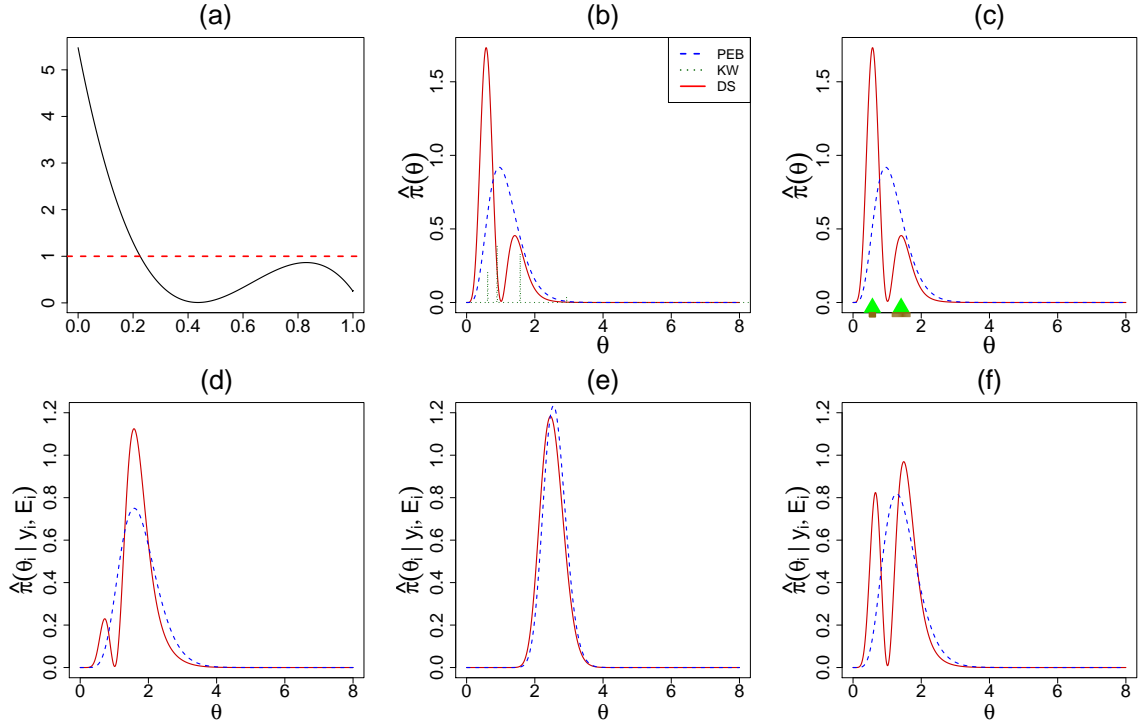


Figure 2.15: Demonstration of DS-Bayes with covariates on the Norberg insurance dataset. Panel (a) displays the U-function. Panel (b) shows the DS-prior (red), the PEB prior (blue) and the Kiefer-Wolfowitz NPMLE prior (green). Panel (c) shows macroinference with standard errors (using smooth bootstrap): two modes located at $0.57(\pm 0.094)$ and $1.41(\pm 0.261)$. Panels (d) through (f) show microinference for occupational groups 13, 22, and 53 (respectively).

2.4.2 The Galaxy Example

The second example consists of $k = 324$ observed rotation velocities y_i and their uncertainties of Low Surface Brightness (LSB) galaxies (De Blok et al., 2001). In addition, we also have the physical radius x_i for each galaxy. We can incorporate the radius into our analysis, resulting in the following DS-model for the galaxy data:

$$\begin{aligned}
 y_i | \theta_i &\sim \mathcal{N}(\theta_i + x_i \nu, \sigma_i^2), \quad i = 1, \dots, 324 \\
 \theta_i &\sim \text{DS}(G, m), \quad G \equiv \mathcal{N}(\mu, \tau^2)
 \end{aligned}$$

The resulting covariate-adjusted the marginal and posterior distributions can be shown to possess the following forms:

$$f_G(y_i; x_i) \sim \mathcal{N}(\mu + x_i\nu, \sigma_i^2 + \tau^2)$$

$$\pi_G(\theta_i | y_i, x_i) \sim \mathcal{N}(\lambda_i\mu + (1 - \lambda_i)(y_i - x_i\nu), (1 - \lambda_i)\sigma_i^2),$$

where $\lambda_i = \frac{\sigma_i^2}{\sigma_i^2 + \tau^2}$. The key observation is we need the residual $y_i - x_i\nu$ to determine the appropriate LP coefficients for our DS(G, m) model. We use MLE to determine starting parameters for g and linear regression to determine ν , then optimize all parameters based on the MLE of the new marginal distribution. Next, we use these parameters to find our residual then apply our algorithm to find the DS(G, m) prior. The estimated DS(G, m) priors for both cases are summarized below:

- (a) Galaxy data (no covariate), with $g(\theta) \sim \mathcal{N}(\hat{\mu} = 85.5, \hat{\tau}^2 = 3300)$:

$$\hat{\pi}(\theta) = g(\theta)[1 + 0.21T_3(\theta; G) - 0.20T_4(\theta; G) + 0.41T_5(\theta; G)] \quad (2.4.2)$$

- (b) Galaxy data (with covariate x_i), with $g(\theta) \sim \mathcal{N}(\hat{\mu} = 51.5, \hat{\tau}^2 = 2050)$ and $\hat{\nu} = 6.2$:

$$\hat{\pi}_{\text{COV}}(\theta) = g(\theta)[1 - 0.30T_2(\theta; G) + 0.24T_3(\theta; G) - 0.28T_4(\theta; G)] \quad (2.4.3)$$

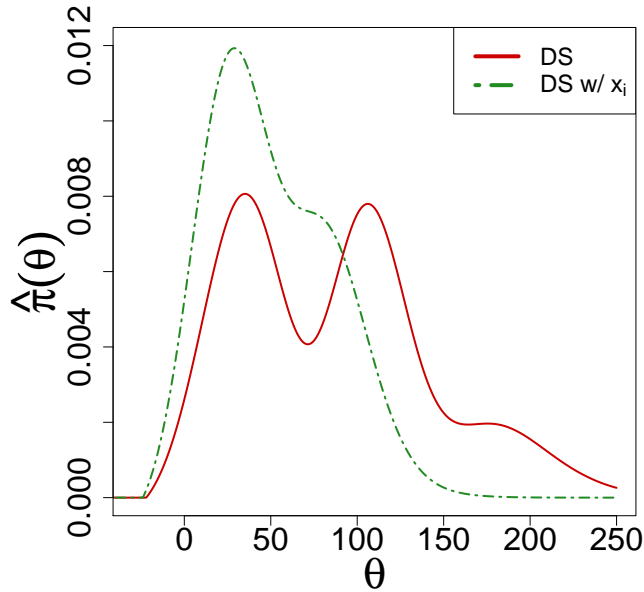


Figure 2.16: The solid red line represents the DS-prior (without covariate) and the dashed green line is the DS-Prior when accounting for the covariate radius x_i .

The solid red line in the Figure 2.16 shows the DS prior for the galaxy data without the radius covariate. We see bimodality among the galaxy velocities, along with a very heavy tail on the far right. Based on macroinference results for the non-covariate model, we confirm two modes: $30.7(\pm 7.74)$ and $108.5(\pm 21.5)$. The dashed green line in Figure 2.16 shows the DS prior that includes the radius covariate x_i . We no longer see bimodality; instead we see a ‘shoulder’ on the right side of the distribution. The inclusion of the covariate **radius** consolidates the two distinct groups of galaxy velocities into one prominent mode around $29.1(\pm 1.84)$.

Through two distinct examples, we showed how the gEB framework can seamlessly incorporate covariates. In the case of the Norburg example, we use the exposure covariate to improve the ability of model to capture the specific risk of a policy holder. In this data set, policy holders who have more exposure to the hazard are more of a risk to make a claim than those who have had less exposure. For the galaxy data set, we see how the incorporation of covariates can alter the prior distribution. Without accounting for the radius, we clearly have two groups of galaxy velocities;

the inclusion of a galaxy’s radius, though, changes it from two to one group. The covariates help inform the model and portray a more accurate story that represents the data. Through both examples, the inclusion of covariates provides a more accurate distributional model that will aid in both inference and interpretation.

2.5 Connections

Now that we have detailed the gEB framework and the $DS(G, m)$ prior, we want to explore the relationship of this approach with other existing Bayesian modeling cultures. By considering these alternative philosophical and computational perspectives, we will show that our formulation is interpretable from many diverse perspectives.

2.5.1 Robust Bayesian Methods

Robust Bayesian methods, outlined in Berger (1994), studies the sensitivity of Bayesian inference as it pertains to the model or prior distribution. Given some random variable $Y \sim f_L(y|\theta)$, θ represents the parameters of interest and the posterior density $f(\theta|y)$, and f_L and f are members of the classes of densities F and Γ . Suppose that $\psi(f, f_L)$ is some inference function (e.g. the mean or mode) for the posterior distribution, then the lower and upper bounds of the inference function $\psi(f, f_L)$ are denoted as

$$\psi_L = \inf_{f_L \in F} \inf_{f \in \Gamma} \psi(f, f_L), \quad \psi_U = \sup_{f_L \in F} \sup_{f \in \Gamma} \psi(f, f_L). \quad (2.5.1)$$

Given the range of possible estimates (ψ_L, ψ_U) , the result is “robust” if this range is small enough to make the conclusion about the inference clear (Berger, 1994). 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 Berger (1994). In the same spirit of the ϵ -contamination class (Berger and Berliner, 1986), our U-function $d(u; G, \Pi)$ can be thought of as an automatic robustifier for standard (conjugate) pri-

ors. Thus, our approach may attain similar goals in a more computationally friendly way. Finally, we completely agree with Berger (1994) 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.’

2.5.2 Empirical Bayes Methods

Neither parametric nor nonparametric, the gEB framework represents a unique blend and compromise between conventional parametric empirical Bayes (PEB) and the Robbins-style full-fledged nonparametric empirical Bayes (NEB). The parametric empirical Bayes (PEB) approach represents a unique tool set to determine an initial g for interrogation in our algorithm. For NEB approaches, we present two examples comparing our approach with empirical Bayes deconvolution and non-parametric maximum likelihood estimates.

Example 1. The dotted line in Figure 2.17(a) denotes the non-parametrically estimated Efron’s $\hat{\pi}$ based on two-dimensional sufficient vector $S = (\theta, \theta^2)$ for the ulcer data (Efron, 1996). 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 2.4(b). 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 nonparametric answer in Efron (1996), -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.

[†]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 alluded to his article ‘Why isn’t everyone a

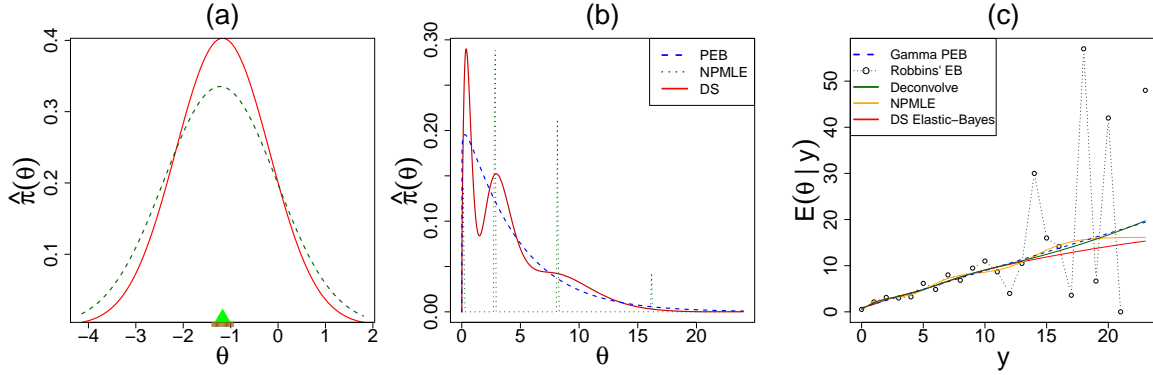


Figure 2.17: Comparisons of $DS(G, m)$ (red) with other empirical Bayes modeling cultures (green): (a) The DS-estimated prior is compared with the exponential prior model from Efron (1996); (b) The DS distribution for the child illness data compared to NPMLE (the dotted line); (c) Estimates for the number of illnesses in the following year $\hat{E}(\theta | x)$ by Gamma PEB, Robbins’ formula, Bayesian deconvolution, NPMLE, and our elastic-Bayes estimate.

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 2.17(b) shows its application to the child illness data (Wang, 2007), 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 $\hat{\pi}(\theta)$ where $g(\theta)$ is a gamma distribution with $\hat{\alpha} = 1.06$ and $\hat{\beta} = 4.19$ as

$$\hat{\pi}(\theta) = \text{Gamma}(\theta; \alpha, \beta) [1 - 0.13T_3(\theta; G) - 0.28T_6(\theta; G)]. \quad (2.5.2)$$

Our method produces a smooth, grid-free $\hat{\pi}$ that accurately captures the overall shape. Figure 2.17(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.

Bayesian?” (Efron, 1986)

Table 2.5: 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™ i5-7200 CPU @ 2.50GHz with Windows 10. DPpackage uses C++ compiler 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

2.5.3 Dirichlet-Process-based Approaches

Bayesian nonparametric [BNP] technique assigns prior distribution on infinite-dimensional spaces of probability models (Ferguson, 1973b). When compared to our algorithm, the computational cost of BNP is severe and produces prior on a set of discrete probability measures that demands an additional layer of smoothing. While BNP is extremely flexible, the heavy cost lies in the task of estimating a massive number of parameters. Contrast this method 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 2.5) and produces smooth estimates in one-shot.

Additionally, Figure 2.18 contrasts Dirichlet-process based Beta-Binomial models (Liu, 1996) with our DS-Bayes model. BNP methods require careful tuning of several hyper-priors values; the result is sensitive to these hyper-prior parameters. 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 also highly complex and require computationally expensive MCMC; the beauty of our approach is that it provides compact analytical expressions that make the computation much more amicable.

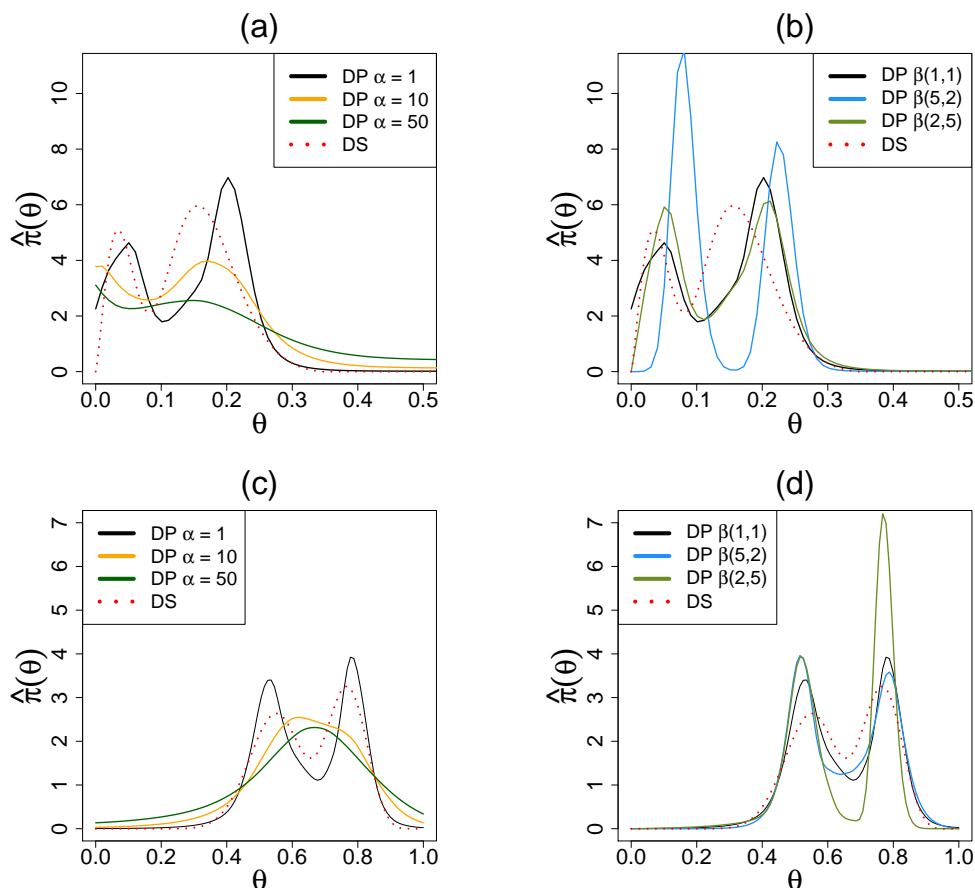


Figure 2.18: Illustrations of the different settings for BNP modeling with a Dirichlet process prior. Panel (a) displays results for the rat tumor data using uniform base prior while varying α . Panel (b), also for the rat tumor data, fixes $\alpha = 1$ and varies the base prior between uniform, Beta(5, 2) and Beta(2, 5). Panels (c) and (d) use the same settings as (a) and (b), but applied to the rolling tacks data.

2.5.4 Weakly Informative Priors

One can also view our approach from a weakly informative prior (WIP) perspective 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” (Gelman et al., 2008). This reformulation

can also bring some tangible computational gain.

2.6 Software

In support of generalized empirical Bayes and the $DS(G, m)$ model, we have developed the R package “BayesGOF” (Mukhopadhyay and Fletcher, 2018a). This package includes functions that implement the $DS(G, m)$ prior and includes all datasets analyzed in this chapter. The appendix provides the R code that demonstrates the package’s capability to execute the estimation, macroinference, and microinference from this chapter.

2.7 Summary

In this chapter, we introduced the theoretical foundation of the generalized empirical Bayes approach with the $DS(G, m)$ family of prior distributions. The core motivation behind our approach is more than just developing another formula for a prior distribution. Instead, we sought to develop a method that can protect analysts and scientists from unqualified specifications of prior distribution. Furthermore, it provides a theory that is developed from a few basic principles and general enough to include commonly-used models. Our algorithm yields an analytic closed-form solution for posterior modeling, noteworthy for the simple reason that none of the nonparametric methods can stand by this claim. Instead of blindly ‘turning the crank’ in the traditional Bayesian manner, gEB encourages an interactive data analysis that often leads to more insights into the data.

Most importantly, gEB represents a third empirical Bayes culture. Our contribution combines the best of both worlds: it will reduce to PEB when the parametric g is appropriate (ulcer and terbinafine data sets) and, in the event of prior-data conflict (rat tumor or child-illness data sets), it automatically produces reliable nonparamet-

ric procedures. This behavior is rooted in the U-function $d(u; G, \Pi)$, which acts as the ‘connector’ between these two extreme philosophies. Our hope is that the generalized empirical Bayes (gEB) framework expedites the development of a new genre of ‘unified’ Bayesian algorithms (Berger, 2000) by leveraging the advantages and benefits of two extreme empirical Bayes philosophies.

CHAPTER 3

UNCERTAINTY MODELING

Modern data sets share three characteristics: heterogeneity, high volume, and low quality (see Figure 3.1) (Lukoianova and Rubin, 2014; Reimer and Madigan, 2018). Heterogeneous data is inconsistent data; they represent measurements and observations of a certain quantity with some statistical and systematic difference (Higgins and Thompson, 2002). High volume is fairly self-explanatory: the data is made up of large quantities of observations. Finally, we define low quality data as data that is noisy and imprecise.

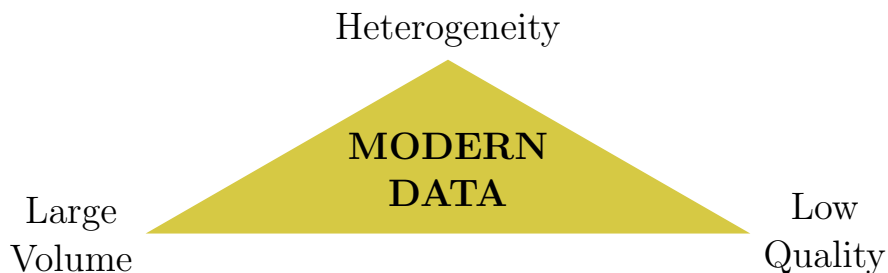


Figure 3.1: Modern data is data that is heterogeneous, large in volume, and low quality.

In an ideal setting, all observed measurements $\{y, s\}$ will consistently and reliably estimate a quantity's unknown true value θ . In reality, the nature of uncertain data

produces measurements y_i such that:

$$y_i = \theta_i + \epsilon \tag{3.0.1}$$

where $\epsilon \sim \mathcal{N}(0, s_i)$. The ϵ_i represents the uncertainty associated with measurement y_i . There are two kinds of uncertainty encapsulated by ϵ_i : statistical and systematic. Statistical uncertainty represents the random fluctuations associated with different measurements; it is typically expected and can be decreased by large sample sizes (Bailey, 2017). On the other hand, systematic uncertainty has origins in either mistakes in measurement or unknown theoretical components (Chen et al., 2003; Bailey, 2017, 2018). Systematic uncertainty is difficult to estimate and contributes to unreliable inference that distorts both data-driven discoveries as well as the reproducibility of scientific research (Ioannidis, 2005; Goodman et al., 2016; Bailey, 2017). How do we extract *reliable* inference from these volumes of *uncertain* data? The key lies in understanding the consistency and reliability of both the measurements y and their uncertainty s (Bailey, 2017, 2018).

3.1 Dealing with Uncertain Data

We propose a four tiered modeling approach to gain reliable inference from uncertain data:

- Tier 1: Find $\hat{\pi}(\theta)$ and provide a combined estimate of the quantity's measurements $\hat{\theta}$ along with its standard error.
- Tier 2: Find $E[\theta | y = y_i] = \hat{y}_i$ and demonstrate the influence of measurement uncertainty s_i on the the observed y_i .
- Tier 3: Given the normalized differences of a measured value z_{ij} , determine the consistency and accuracy of reported uncertainties $f(z; Z)$ en quantity.

- Tier 4: Given $\hat{\pi}(\theta)$ and a new measurement y^* , find θ^* and assess the quality of s^* .

3.1.1 Tier 1

Since the purpose of any scientific study is to estimate the true value of a quantity θ through deliberate and precise measurements y , the first tier aims to understand how uncertainty in the observed y influences our understanding of θ . For many quantities, multiple studies have produced varied and distinct measurements; some generate very exciting results that ultimately turn out to be incorrect (Bailey, 2017). Furthermore, each study may take a different approach to estimating θ and thus may not share the same systematic uncertainty as other inquiries. Given these factors, we can view each study's measured value y as in (3.0.1) where θ is governed by a distribution $\pi(\theta)$. The estimate of this distribution $\hat{\pi}(\theta)$ and its functional shape are key to finding $\hat{\theta}$. Not only will $\hat{\pi}$ help us find $\hat{\theta}$, but it also describes all the shared uncertainty from the various studies.

Given previous k studies into a quantity, each study reported observed measurements y_i and its corresponding uncertainty s_i , with $i = 1 \dots k$. Our scenario leads to the following model:

$$y_i | \theta_i \stackrel{\text{ind}}{\sim} \text{Normal}(\theta_i, s_i), \quad (i = 1, \dots, k)$$

$$\Theta_i \stackrel{\text{ind}}{\sim} \pi(\theta),$$

When $\pi(\theta) \sim \text{Normal}(\mu, \tau^2)$, we have the standard random effects meta-analysis model that allows us to estimate $\hat{\theta}$ using the mean μ . What if $\pi(\theta)$ is not normally distributed? In this situation, we cannot use the mean to estimate θ and should seek ways to correct our estimate (Jeffreys, 1938). Generalized Empirical Bayes and the DS(G, m) model are ideally suited for such a situation. Particularly, the DS(G, m)

prior will determine if the systematic uncertainty of observed measurements y is appropriately described by a Gaussian distribution. We can easily adapt the aforementioned random effects model to one that leverages the $\text{DS}(G, m)$ prior:

$$y_i | \theta_i \stackrel{\text{ind}}{\sim} \text{Normal}(y_i | \theta_i), \quad i = 1, \dots, k$$

$$\Theta \sim \pi(\theta) = \text{Normal}(\mu, \tau^2) \times d[G(\theta)], \quad (3.1.1)$$

where $G(\theta)$ is the distribution function for $g(\theta) = \text{Normal}(\mu, \tau^2)$. When $d[G(\theta)] = 1$, we have the standard random effects meta-analysis where the structure of the heterogeneity only requires the assumed Normal distribution. In situations where $d[G(\theta)] \neq 1$, the uncertainty and heterogeneity among the measurements dictates a different structure for $\pi(\theta)$. Once we determine the proper structure, we can proceed with finding $\hat{\theta}$. Through the $\text{DS}(G, m)$ model, our first tier finds and appropriate $\hat{\pi}(\theta)$ and then subsequently perform macroinference to estimate $\hat{\theta}$ and its standard error.

3.1.2 Tier 2

The second tier of analysis builds on the first. With $\hat{\pi}(\theta)$, we see how systematic uncertainty influences our observations of θ . Microinference uses elastic-Bayes to refine an observed y based on our knowledge of $\hat{\pi}(\theta)$. In terms of uncertain data, we are ‘cleaning’ our measurements of any shared systematic uncertainty demonstrated in $\hat{\pi}(\theta)$. The final result, $\mathbb{E}[\theta_i | y = y_i]$, is an estimate of θ_i that includes systematic uncertainty unique to that particular y_i .

3.1.3 Tier 3

The third tier of our analysis looks at the consistency and accuracy of reported uncertainties. The conventional assumption, given *quality* data, is that the error distribution of a quantity should be Gaussian (Crandall et al., 2015; Bailey, 2017, 2018). As

this assumption drives our understanding of what constitutes a data-driven discovery, the goal of this tier is to determine the true distribution of uncertainty $f(z; Z)$. We find $f(z; Z)$ using the normalized differences between measured values z_{ij} (Bailey, 2018) with one of two approaches. Both approaches model the error distribution in a manner that gives insight into how systematic uncertainty is distributed throughout a quantities measurements (Jeffreys, 1938; Bailey, 2018).

The first approach to find z_{ij} seeks to determine if all observed y are consistent with their reported uncertainties s (Bailey, 2017, 2018). Let z_{ij} be the the normalized difference for each pair of observed y_i and s_i in the studied quantity:

$$z_{ij} = \frac{(y_i - y_j)}{\sqrt{s_i^2 + s_j^2}} \quad (3.1.2)$$

The second approach uses an established value for a quantity as a basis of comparison for all observed y . Let y_* be the established value, then

$$z_{i*} = \frac{y_i - y_*}{s_i} \quad (3.1.3)$$

Although a subtle difference, using y_* provides insight into the compatibility of the established value with other results (Chen et al., 2003; Bailey, 2017).

When analyzing the z_{ij} 's, there is no underlying assumption of how the z_{ij} 's are distributed. These observations are 'unconditional' and are not governed by a prior distribution. In our development of the algorithm for the DS prior (Chapter 2.2.2), we pointed out that the $\theta_i, \dots, \theta_k$ are unobserved thus requiring an application of the law of iterated expectations to generate 'ghost' LP mean coefficients. In the case of the z_{ij} , we have the true observations and do not require any conditional expectations. While gEB is inappropriate, we can still apply the DS(G, m) model to identify discrepancies between the assumed normality and reality.

We can determine the LP mean coefficients as in Mukhopadhyay (2017). From

Definition 1, we have the Skew-G class of density models as given by

$$f(z; Z) = g(z; \mu, \tau^2) d[G(z); G, F],$$

where $d(u; G, F) = \frac{f(G^{-1}(u))}{g(G^{-1}(u))}$ for $0 < u < 1$ and consequently $\int_0^1 d(u; G, F) = 1$. Given the same construction of the LP basis functions $T_j(Z; G)$, the LP mean coefficients are

$$\text{LP}(j; G, F) = \mathbb{E}[T_j(Z; G) \mid F] \quad (3.1.4)$$

Although its derivation is unique from the DS prior, the resulting DS distribution $f(z; Z)$ provides the same insights as one found using gEB. We still have the uncertainty function $d[G(z); G, F]$ to diagnose any corrections to the assumed distribution g . Furthermore, we can apply the same smoothing criteria to the LP means of $f(z; Z)$ as we do to $\pi(\theta)$. The uniqueness of $f(z; Z)$ is that we are no longer constrained by a conjugate prior g . Without the conditional restraints of a prior distribution, we can now select any appropriate distribution for g . In terms of the uncertainty modeling, we will follow Bailey (2017) and use the t-distribution with 2 degrees of freedom for our analysis.

For demonstration purposes, we explore data sets introduced in Bailey (2017). The author showed that the distributions of uncertainty for studies in the fields of medical and particle physics research requires heavier tails than a Normal distribution; in both cases, the author found that a non-standardized Student's t-distribution with 2 or 3 degrees of freedom and their heavy tails more appropriate for these data sets. By selecting $g(z)$ as the t-distribution with 2 degrees of freedom, we can validate $g(z)$ is appropriate for the pairwise data. The results are shown in Figures 3.2.

First, we discuss the particle physics uncertainty data. From the comparison density function in Figure 3.2(a), we see minor adjustments that slightly deviate from uniform. This behavior is our first indication that our final $\hat{f}(z)$ will not vary

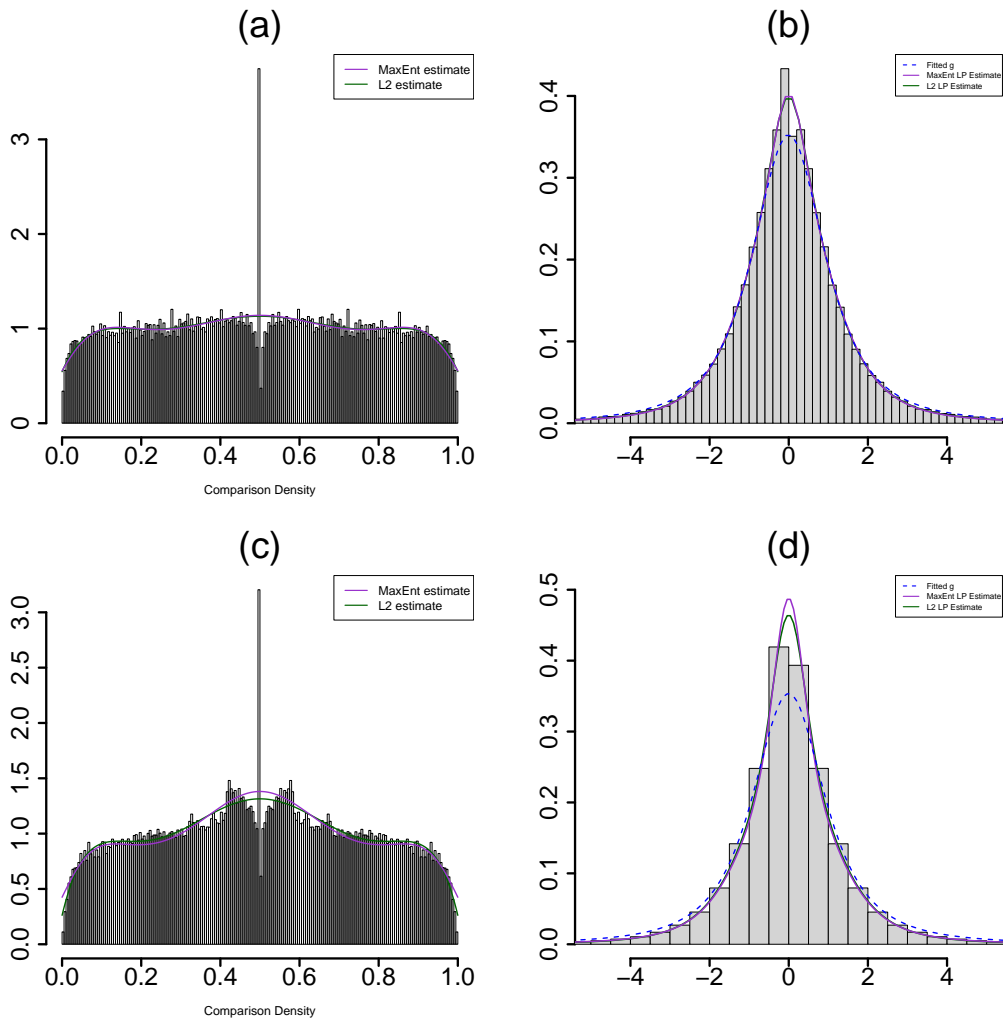


Figure 3.2: Tier 3 analysis for the physics and medical uncertainty data from Bailey (2017). Panel (a) and (c) shows the comparison density function overlaid on a histogram of the quantile values for the physics and medical uncertainty data, respectively. Panels (b) and (d) depict the original g (blue dashed), along with the maximum entropy (purple) and L2 (green) representation of the DS distribution.

much from g . The equation for our $DS(G, m)$ distribution $f(z)$ is

$$\hat{f}(z; Z) = g(z; 2)[1 - 0.199T_2(z; G) - 0.081T_6(z; G)]. \quad (3.1.5)$$

and displayed in Panel (b) of Figure 3.2. While (3.1.5) includes $T_2(z; G)$ and $T_6(z; G)$, the resulting changes to the variance and behavior at the tails do not indicate a significant difference from the proposed $g(z)$: the t-distribution proposed by Bailey (2017). The medical uncertainty data shows a similar result as the particle physics analysis. The comparison density shows some adjustment to the tails. The equation for $\hat{f}(z; Z)$ dictates a small adjustment to the variance of $g(z)$ through the significant $T_2(z; G)$:

$$\hat{f}(z; Z) = g(z; 2)[1 - 0.199T_2(z; G)]. \quad (3.1.6)$$

From Figure 3.2(d), we see no indication that there is much difference from the original g . Through our demonstration of tier three analysis, we demonstrated how the $DS(G, m)$ model assesses both the consistency and accuracy of reported uncertainties. For this example, we confirmed and reproduced the result in Bailey (2017).

3.1.4 Tier 4

Suppose we have a new measurement y^* for a quantity and its corresponding uncertainty s^* . The quantity in question has been previously explored, with k established measurements. Can we use the established measurements for the quantity to improve y^* ? Furthermore, can I use those previous measurements to assess my uncertainty s^* ? The fourth tier addresses both questions using established techniques for gEB.

First, we apply microinference (Chapter 2.3.3) on (y^*, s^*) given our established $\hat{\pi}$ from the first tier. The result is a posterior distribution $\hat{\pi}(\theta | y = y^*)$ and subsequently a posterior mean $\mathbb{E}[\theta | y = y^*]$. The posterior mean represents the expectation of the true measurement θ given our new observed measurement y^* . This posterior mean

utilizes elastic Bayes from (2.3.2) to generate the improved assessment for y^* .

The next step of this tier is to assess the uncertainty estimation of s^* . The goal is to generate an estimate of a posterior density that incorporates variability in the hyperparameters, which in this case would be μ and τ^2 . From this estimate, we provide a credible interval for y^* that serves as a diagnostic measure for s^* . This assessment utilizes a technique called Finite Bayes inference, which integrates microinference with the accept / reject sampling algorithm utilized in macroinference. While estimating prior parameters based on the observed data is predominant in parametric empirical Bayes, those that follow a more traditional Bayes approach see this method as problematic and detrimental to the final inference. Traditional Bayesians argue that ‘double dipping’ with the data (estimating both the prior and posterior with the same data) ignores some posterior uncertainty and overestimates the precision (Gelman et al., 2013). To avoid these problems, the traditionalists encourage the use of a hyperprior $h(g)$ for the parameters of the selected $g(\theta)$. The resulting model has the form

$$\begin{aligned} y_i | \theta_i &\stackrel{\text{ind}}{\sim} f(y_i | \theta_i), & (i = 1, \dots, k) \\ \Theta_i &\stackrel{\text{ind}}{\sim} \pi(\theta | \phi), \\ \Phi &\stackrel{\text{ind}}{\sim} h(\phi), \end{aligned}$$

The hyperprior h introduces uncertainty with the prior’s parameters, thus preserves the observed data for use only in posterior analysis. How do we appease these traditional Bayesians and leverage the uncertainty of hyperpriors without sacrificing a scientific g ?

The answer lies in finite Bayes inference discussed in Efron (2017). The primary goal of finite Bayes inference is to incorporate the variability of our estimated prior parameters $\hat{\mu}$ and $\hat{\tau}^2$ into our posterior estimates. While Morris (1983) introduced

this concept for a Gaussian prior, Efron (2017) expanded it for any prior distribution g . We can use Efron’s adaptation along with the $DS(G, m)$ sampling algorithm in Section 2.3 to extend finite Bayes inference to the family of $DS(G, m)$ priors:

gEB Finite Bayes Algorithm

Step 1. Estimate $DS(G, m)$ prior $\hat{\pi}(\theta)$ using observed data y_i , where $i = 1, \dots, k$ and selected g .

Step 2. Generate bootstrap data set $\mathbf{y}^* = (y_1^*, \dots, y_k^*)$ using $DS(G, m)$ sampling.

Step 3. Use \mathbf{y}^* to estimate $\hat{\pi}$.

Step 4. Simulate a large number N of bootstrap posteriors $\hat{\pi}$ and average for a corrected prior,

$$\tilde{\pi}^*(\theta) = \frac{1}{N} \sum_{j=1}^N \hat{\pi}^{*j}(\theta)$$

Step 5. Conduct microinference using $\tilde{\pi}^*(\theta)$ and new observation y_0 .

By using a bootstrap distribution for $\tilde{\pi}^*(\theta)$, we can replicate the uncertainty of using a non-informative hyperprior $h(g)$ on the the posterior variability. Each of the N bootstrap posteriors are generated from \mathbf{y}^* , which in turn will have their own unique parameters for g . It’s the variation of these parameters among the N posterior estimates that replicates the variability introduced with a hyperprior. More importantly, the algorithm provides a route to validate our point estimates for our estimate prior parameters through bootstrapping parametric data sets from the marginal density. This process will replicate the variability and heterogeneity of uncertain data and allow us to assess the measured uncertainty s with a credible interval of the posterior distribution.

3.2 Real Data Applications of Tiered Analysis

We will apply our tiered analysis to three data sets: the Hubble Constant (H_0), Newton’s gravitational constant (G_N), and the Lithium abundance measurements (${}^7\text{Li}$). The Hubble constant is an estimate of the rate of the universe’s expansion. Once the value of H_0 was the source of much controversy, it is now one of the more precisely measured cosmological parameters (Chen et al., 2003). The data set includes $k = 477$ measurements from 1927 through 2003 (Huchra, 2008). Newton’s gravitation constant, known as “Big G”, determines the strength of gravity. While it has been measured by scientists for over two hundred years, G_N has a long history of producing inconsistent measurements (Bailey, 2018). The data set includes $k = 48$ measurements from 1798 through 2013. Finally, the Lithium abundance measurements are linked to another cosmological parameter used in standard Big Bang nucleosynthesis and includes $k = 66$ observations. ${}^7\text{Li}$ is very fragile and found in the atmospheres of stars (Crandall et al., 2015).

3.2.1 Hubble Constant H_0

First, we address the data for the Hubble Constant. For first tier analysis, we assume a random effects model and use method of moments to determine the parameters for $g \sim \text{Normal}(67.5, (13.3)^2)$. Next, we estimated $\hat{\pi}(\theta)$ and subsequently, our ‘clean’ θ estimate. Figure 3.3 shows the corresponding uncertainty function and macroinference.

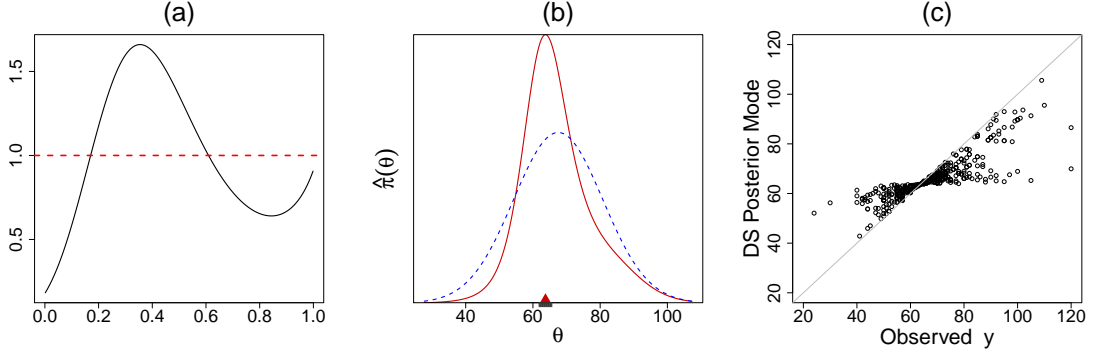


Figure 3.3: For the Hubble data set, Panel (a) shows the uncertainty function which indicates structure change for the right tail, while (b) details $\hat{\theta}$ with a single mode at 63.7 ± 1.52 . Panel (c) shows the Tier 2 analysis, with DS-elastic Bayes shrinking toward the mode.

From the uncertainty function in Figure 3.3(a), $\pi(\theta)$ requires some adjustments for the mode and the right side tail. Figure 3.3(b) shows $\hat{\pi}$ compared to the original assumed $g(\theta)$. The \mathcal{L}^2 equation for $\hat{\pi}(\theta)$ is:

$$\hat{\pi}(\theta) = g(\theta; \mu, \tau^2) [1 - 0.30T_2(\theta; G) + 0.25T_3(\theta; G)]. \quad (3.2.1)$$

The significant LP means at T_2 and T_3 reflect the corrections to $g(\theta)$: a slight decrease to the variability of the distribution and increasing its positive skewness. The macroinference on the mode provides $\hat{\theta} = 63.7 \pm 1.52$.

Figure 3.3(c) shows the results of Tier 2 analysis. Here, DS-elastic Bayes shrinks the observed y to the mode of $\hat{\pi}(\theta)$. The larger spread of the values to the right of the mode depicts the influence of the corrected $DS(G, m)$ prior on the shrinkage.

We will conduct the third tier of analysis using a comparison value. We calculate z_{ij} using $H_0 = 71$ for our central value, which was determined by Hinshaw et al. (2003) and used by Chen et al. (2003). Using this comparison value, Chen et al. (2003) found the z_{ij} values followed a modified $t(2)$ distribution; therefore, we will let $g(z) \sim t(2)$. Figure 3.4(a) and (b) show the comparison density and $\hat{f}(z; Z)$, respectively.

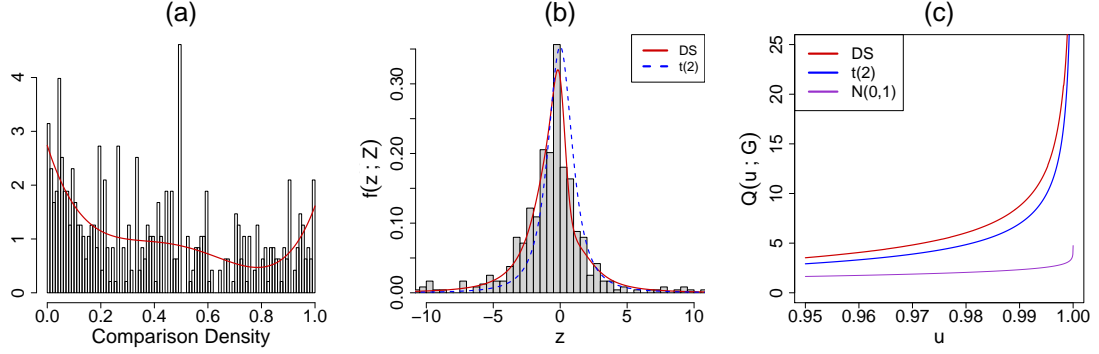


Figure 3.4: Tier 3 analysis for the Hubble constant data set. Panel (a) shows the comparison density. Panel (b) shows the histogram of the observed data along with the $\hat{f}(z; Z)$. Panel (c) displays the quantile as a function of $F(z)$.

The comparison density in Figure 3.4(a) shows that $t(2)$ is not appropriate; the true $f(z; Z)$ requires adjustment to the mean and tails. The equation for the \mathcal{L}^2 representation is

$$\hat{f}(z; Z) = g(z; 2) [1 - 0.32T_1(z; G) + 0.29T_2(z; G) + 0.18T_4(z; G)]. \quad (3.2.2)$$

The significant LP means confirms our intuition from the comparison density. The $T_1(z; G)$ shifts the entire distribution to the left; $T_2(z; G)$ slightly increases the width (and hence the variability); and $T_4(z; G)$ indicates that the tails require slightly heavier behavior than the proposal distribution $g(z; 2)$. Despite the non-uniform comparison density for $g(z) \sim t(2)$, we believe our results support Chen et al. (2003); the authors incorporated a scaling to improve the fit of a $t(2)$ to the uncertainties. Our result shows how $DS(G, m)$ generates a similar adjustment, but without scaling the data. Figure 3.4(c) shows the influence of the $f(z; Z)$ on the quantile function $Q(u; G)$; the red line shows slightly heavier tails than the original $g(z) \sim t(2)$. Table 3.1 details a comparison of the differences.

Table 3.1: Quantile values for the Hubble data set.

u	0.95	0.975	0.99	0.9999
Normal(0, 1)	1.645	1.960	2.326	3.7190
$t(2)$	2.9200	4.3027	6.9646	70.7001
DS	3.5286	5.3395	8.7393	64.7132

For the fourth tier of analysis, we will use the central value from Chen et al. (2003) (based on the work in (Hinshaw et al., 2003)) that $H_0 = 71 \pm 3.5$ (where the standard error is the mean of the upper and lower uncertainty). With $y^* = 71$ and $s^* = 3.5$, the results of Tier 4 analysis are shown in Figure 3.5.

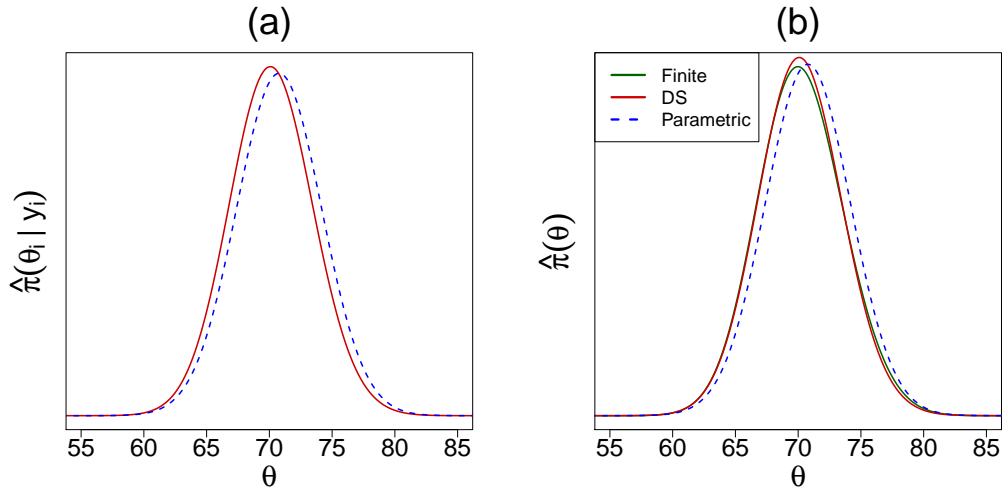


Figure 3.5: Tier 4 analysis of the Hubble constant data set. Panel (a) shows the posterior distribution of $\hat{\pi}(\theta|y^*)$ where the posterior mean is 70.18 and the posterior mode is 70.11. Panel (b) is the Finite Bayes analysis, with mean of 70.22 and mode of 69.95. The 90% credible interval is (66.76, 73.62).

Figure 3.5(a) gives the posterior distribution of $y^* = 71$ where the red represents the posterior distribution based on $\hat{\pi}(\theta)$ and the blue is based on the standard random effects model. The posterior mean $\mathbb{E}[\theta | y = 71]$ is 70.18 and posterior mode is 70.11, indicating that the ‘cleaned’ value θ^* slightly differs from the actual measurement. Figure 3.5(b) shows the second part of the analysis, where the green curve represents

the finite Bayes inference. From this distribution, we determine the 90% credible interval as (66.76, 73.62). In comparison, the same interval based on the observed s^* is (67.5, 74.5) thus our analysis shows the reported uncertainty is appropriate for the observed measurement.

3.2.2 Newton’s Gravitational Constant G_N

First, we analyze “Big G” using two representations of the data set. The first representation includes all $k = 48$ observations. The second, though, includes only $k = 47$. From Bailey (2018), we find that a measurement from 1996 is 65 standard deviations *different* than the 2014 value CODATA established as true. It took scientists eight years to determine that the 1996 measurement was heavily influenced by overlooked variations from a new component used in the study. With such a large difference in the observed measurements, we seek to understand how such a unique observation can influence our understanding of θ .

For the first tier, we assume that the observed measurements for G_N follow the random effects model in (3.1.1). We use a method of moments approach to determine the parameters for g . When $k = 48$, let $g \sim \mathcal{N}(6.67, (0.0035)^2)$. For $k = 47$, we find $g \sim \mathcal{N}(6.67, (0.0012)^2)$. While both cases have the same μ , we find they both also have a very small τ^2 relative to μ . It is important to test if the data follow a fixed effect model, meaning that there is no heterogeneity between the different studies.

We will use both Cochran’s Q test and I^2 to evaluate the heterogeneity of the data. In both $k = 48$ and $k = 47$, the tests gave no evidence of observed heterogeneity among the studies (p-value equal to 1 and $I^2 = 0$). Therefore, we find our estimate of θ using the fixed effect weighted mean: $k = 48$ results in $\hat{\theta} = 6.6742 \pm 0.00003$ while $k = 47$ finds $\hat{\theta} = 6.6741 \pm 0.00003$. The very small difference in $\hat{\theta}$ indicates the large outlier has little impact on our first tier analysis.

Since we pursued a fixed effect model for our first tier, we will bypass the second

tier and move directly to the third. In this tier, we want to compare the distributions of the measurement uncertainty $f(z; Z)$ for both groups of data. For this analysis, we will compute the pairwise difference between all measurements to determine the level of systematic uncertainty shared between the measurements. Following Bailey (2018), we let $g(z) \sim t(2)$ and apply the $DS(G, m)$ model to the Z_{ij} values. The results are shown in Figure 3.6.

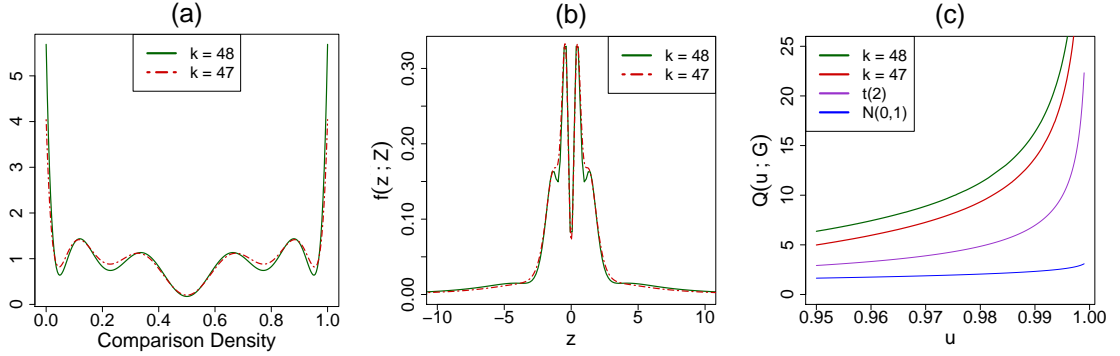


Figure 3.6: Third tier analysis of Big G’s pairwise Z_{ij} with g as the t -distribution with 2 degrees of freedom and $m = 10$. Panel (a) shows the comparison densities for all pairs using $k = 48$ and $k = 47$. Panel (b) compares the DS distribution of $k = 48$ and $k = 47$, where the later shows only two significant modes. Panel (c) shows the difference in the quantile values for each distribution.

Figure 3.6(a) displays the comparison densities for both $k = 48$ (solid green) and $k = 47$ (dashed red). The multimodality of the comparison densities for both data sets indicates significant changes to g . Despite similar behavior between the two comparison densities, $k = 48$ does show more extreme behavior at the tails and sharper dips at 0.2 and 0.8. That behavior translates to the plots of $\hat{f}(z)$ in Figure 3.6(b). The following are the \mathcal{L}^2 ($m = 10$) representations for $\hat{f}(z; Z)$:

- For $k = 48$ (solid green curve):

$$\begin{aligned} \hat{f}_{48}(z; Z) = g(z; 2) & [1 + 0.36T_2(z; G) + 0.12T_4(z; G) + 0.31T_6(z; G) + \\ & + 0.18T_8(z; G) + 0.37T_{10}(z; G)]. \end{aligned} \quad (3.2.3)$$

- For $k = 47$ (dashed red curve):

$$\hat{f}_{47}(z; Z) = g(z; 2)[1 + 0.30T_2(z; G) + 0.22T_6(z; G) + 0.08T_8(z; G) + 0.27T_{10}(z; G)]. \quad (3.2.4)$$

The subtle differences in the comparison density present themselves in the equations for $\hat{f}(z)$. While both had $m = 10$, \hat{f}_{48} has five significant LP coefficients compared to four in \hat{f}_{47} . This difference results in the two distinct functions in Figure 3.6(b) with \hat{f}_{48} having *six* significant modes compared to only two in \hat{f}_{47} . The six modes illustrate additional types of systematic uncertainty introduced by including the extreme 65σ measurement in the analysis.

The final panel, Figure 3.6(c), illustrates the impact of uncertainty on data-driven discoveries, where $u = F(z)$ and $Q(u; G)$ is the associated quantile value. The blue curve represents the perfect case where the uncertainty distribution is $\mathcal{N}(0, 1)$; it shows that as u gets closer to 1, there is almost no change to the corresponding quantile. Under this assumption of normality, one may be confident that they have made a discovery for anything larger than $Q(u; G) = 2$. In contrast, the purple curve represents the t distribution with 2 degrees of freedom; its position with respect to the $\mathcal{N}(0, 1)$ indicate larger probability in the tails. Here, we see how an incorrect assumption about the uncertainty distribution can lead to false discoveries: a significant find under the assumption of normality is not necessarily significant when heavy tails are present (Bailey, 2017).

The curves for both $k = 47$ (red) and $k = 48$ (green) illustrate an even more extreme situation than $t(2)$. In both cases, we do not see 95% of the probability in the distribution until $Q(u; G)$ is at least five. With the inclusion of the outlier in $k = 48$, we require a slightly higher $Q(u; G)$ to cover 95%. Table 3.2 provides a sample of the differences. This identification of significantly heavier tails in the Big

G data set allows scientists to better understand when a measurement qualifies as a true discovery.

Table 3.2: Quantile values for the Big G data set.

u	0.95	0.975	0.99	0.9999
Normal(0, 1)	1.645	1.960	2.326	3.7190
$t(2)$	2.9200	4.3027	6.9646	70.7001
$k = 47$	4.9866	8.1637	13.7247	129.8299
$k = 48$	6.3658	9.9218	16.3624	129.3001

We take a moment to analyze the impact of time on the measurements for G_N . Since these measurements span over 200 years, we expect that the consistency and compatibility of these measurements are related to the methods and tools popular within a given time period. The earliest measurement, from 1798, will not have access to the innovations and devices available to the measurements of the 21st century. By analyzing uncertainty over time, we gain more insight into how innovation impacts measurement uncertainty. We begin by using the three modes from the \mathcal{L}^2 representation in Figure 3.6(b) as the initial centers for a k-means clustering of the $|Z_{ij}|$ values. Table 3.3 shows the original modes identified in $\hat{f}(z; Z)$ and the k-means calculated centers.

Table 3.3: Comparison of original modes identified from $\hat{f}(z)$ to the refined clusters identified by applying k-means to the $|z_{ij}|$.

Original Mode	Refined Cluster	Membership
0.375	1.19	1932
1.28	8.48	286
4.13	57.81	38

Figure 3.7(a) shows the $|Z_{ij}|$ values based on the year of the observation and colored according to their center membership. The black points represent those members of the first cluster, 1.19. The congregation of the black points along the bottom of the graph is not surprising, but it is interesting that each study has $|Z_{ij}|$ within that group; thus, we see some level of consistency among the measurements throughout the 200 year span. The red points correspond to the members of the second cluster at 8.48 and where we begin to see the manifestation of larger inconsistencies with respect to the uncertainties.

It is not until the 1974 study, though, where member's of the final cluster (57.81) begin to appear. The 38 $|Z_{ij}|$'s represent those measurements that are largely inconsistent with others; the majority appear to congregate around 1995 and 1996. Figure 3.7(b) shows that the 1996 measurements are responsible for all the members of the largest cluster. This image maintains a similar color scheme as panel (a), but the black represents no group and the grey are those that members of the first cluster. These measurements include the one Bailey (2018) identified as 65 standard deviations different from a recent measurement.

Figure 3.8(a) shows the results of fourth tier analysis for a new measurement in the Big G data: the current official 2014 CODATA value of 6.67408 ± 0.00031 . From our microinference, we found $\mathbb{E}[\theta | y = 6.67] = 6.6741$. While not significantly different than the original measurement, the slight difference results from the influence of how $\hat{\pi}(\theta)$ models the underlying uncertainty of θ_i .

Panel (b) shows the application of Finite Bayes to the Big G data and the 2014 CODATA value. We generated the estimate $\mathbb{E}[\theta | y = 6.67] = 6.6742$ and derived a 90% Bayesian credible interval for y^* as $(6.6736, 6.6746)$. In order to best compare with the provided standard error, we also generated a 68% credible interval for $\hat{\pi}(\theta | y = y^*)$. The final result, $(6.6738, 6.6744)$, equates to an approximate $\hat{s}^* = 0.0003$ compared to the measured uncertainty of $s^* = 0.00331$. In this case, we can say that

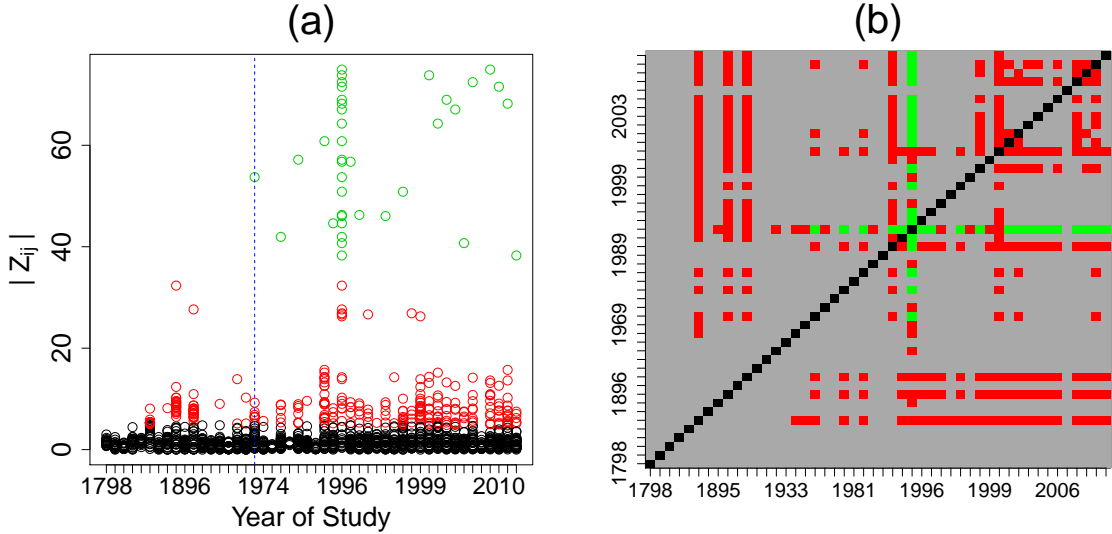


Figure 3.7: Panel (a) shows the pairwise comparisons of each study, where the colors reflect the groups identified through application of k-means clustering with the positive modes in Figure 3.6(b). The blue dashed line represents the first point where a measurement contains a value associated with the largest mode. Panel (b) shows a matrix plot of the groupings. The green portion of the plot represents the pairwise differences that are most extreme and coincide with the 1995-1996 measurements.

the reported uncertainty may be slightly inflated for s^* .

3.2.3 Lithium Abundance ${}^7\text{Li}$

Finally, we look at the ${}^7\text{Li}$ data set using tiers one through three. We begin our tier one analysis by using method of moments to find $g \sim \text{Normal}(2.21, (0.044)^2)$. After applying generalized Empirical Bayes modeling to determine $\hat{\pi}(\theta)$, we find g is appropriate distribution for the random effects model for the ${}^7\text{Li}$ measurements. Figure 3.9(a) shows the macroinference, which results $\hat{\theta} = 2.21 \pm 0.009$. In comparison, Crandall et al. (2015) used a weighted mean to determine a central estimate of $\hat{\theta} = 2.20 \pm 0.0044$.

Figure 3.9(b) shows the tier two analysis. Since $\hat{\pi}(\theta) = g(\theta)$, we find $\mathbb{E}[\theta_i | y = y_i]$ through a standard application of Stein’s shrinkage. The ‘bowtie’ shape, similar to what we saw with the Hubble constant, indicates how the observed y are pulled

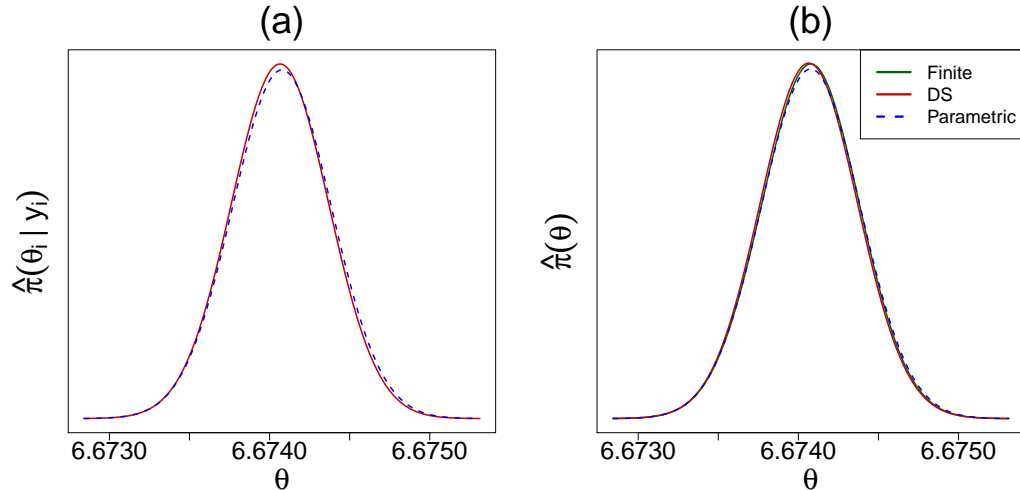


Figure 3.8: Tier 4 analysis of the Big G data set. Panel (a) shows the posterior distribution for $y^* = 6.67408$. The posterior mean and mode are both 6.6741. Panel (b) shows the finite Bayes inference, with mean of 6.6742 and mode of 6.6741. The 90% credible interval is (6.6738, 6.6744)

toward the mean of $g(\theta)$. Essentially, this analysis is ‘cleaning’ the observed y_i of the shared systematic uncertainty.

Finally, we use tier three analysis and generate the z_{ij} based on the median central estimate of 2.21. From this estimate, Crandall et al. (2015) determined that the uncertainty distribution was better described by a t-distribution with 2 degrees of freedom, thus we begin with $g(z) \sim t(2)$. As with our Tier 1 analysis, our Tier 3 analysis finds $\hat{f}(z; Z) = g(z)$. Therefore, we have confirmed the results from Crandall et al. (2015) the uncertainty distribution is not Gaussian.

3.3 Summary

This chapter applied our four-tiered approach to modern data analysis to three unique examples, resulting in improved inference for θ and a better understanding of how uncertainty influences these inferences. We showed how the first tier uses macroinference to find a consensus value for the true value θ . The second tier applies microinference to clean and ‘purify’ observed measurements muddled by uncertainty, giving researchers

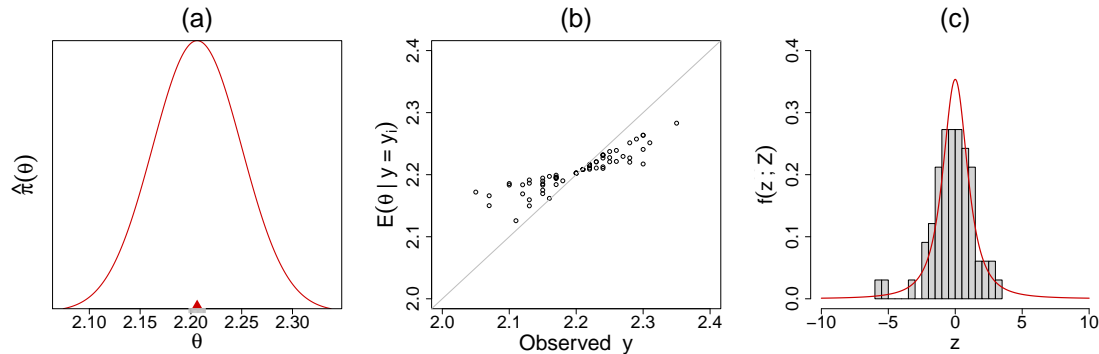


Figure 3.9: Tier 1 through 3 analysis of the ${}^7\text{Li}$ data. For Tier 1, $\hat{\pi} \equiv g$ thus Panel (a) shows one significant mode at 2.21 ± 0.009 . Panel (b) provides the Tier 2 analysis, which is an application of Stein’s shrinkage. Panel (c) shows $f(z; Z)$ using comparisons to the median value. As with Tier 1, the Z_{ij} require no correction and thus $f(z; Z) \equiv t(2)$

better insight into their findings. The third tier of analysis is key in determining the appropriate level of significance for a quantity, particularly when the true distribution of uncertainty has significantly different behavior at the tails than the standard Gaussian assumption. Finally, the fourth tier incorporates a finite Bayesian approach with microinference to estimate a new measurement’s θ_i and evaluate its uncertainty.

While all four-tiers will aid scientists and researchers in their work with uncertain data, perhaps the most insight will come from the third tier. The high uncertainty associated with modern data creates situations that can confound data driven discovery, particularly with regard to what constitutes a significant measurement or finding. This criteria is critical to establishing reliable new discoveries and our analysis shows how uncertain data makes the use of the ‘school-house’ answer of 5% (or two deviations from the standard Normal) laughable. While several fields have increased the standard of discovery to five, or even seven, deviations (Chen et al., 2003; Bailey, 2017), our third tier of analysis shows that even these concrete control measures do not ensure that an observed result is both significant and accurate.

Finally, we want to highlight how the first tier of analysis relates to meta-analysis. The structure of the model, a normal likelihood with a normal prior distribtuion, is

the same as what is found in random effects meta-analysis (Hedges, 1983). With the addition of the $DS(G, m)$ prior into the model, we have introduced a new form of non-parametric random-effects meta-analysis. The uncertainty function d serves as the measure of heterogeneity between the studies. When $d = 1$, then the Normal distribution is appropriate for the random effects model. In the event $d \neq 1$, we find that the heterogeneity between the studies requires more than a simple Normal; not only does d provide a visual diagnostic of the heterogeneity, but we can also quantify it through qLP as discussed in Chapter 2.

CHAPTER 4

ADDITIONAL APPLICATIONS

In Chapters 2 and 3, we demonstrated its effectiveness when applied to a variety of problems based on hierarchical models. We provided a glimpse of how we can expand gEB's reach in Chapter 3, by extending the $DS(G, m)$ prior to the DS distribution. This extension effectively developed a method to model unconditional data without dependence on a hierarchical structure. In many of the previous applications, we found new insights due to gEB's ability to handle heterogeneity. Now, we present two unique applications that will further broaden the spectrum of utility for gEB and the $DS(G, m)$ prior.

The first application focuses on gaining inference from Big Data. In particular, we are going to illustrate how gEB and the $DS(G, m)$ prior can provide a more systematic approach to predictive modeling from heterogeneous distributed architecture. The second application is the missing species problem, which is a well-studied problem with numerous approaches to its solution. Whereas many of these techniques are tailor-made for this specific problem, we will use the missing species problem to demonstrate how we can use gEB to solve this problem with two distinct approaches.

4.1 Inference from Distributed Data

Although relatively new, Big Data has become a permanent fixture of business, government, and society. Before the advent of Big Data, statistical inference was critical in quantifying the uncertainty of an estimator $\hat{\theta}$ for some population parameter θ . Prior to modern computing and storage capabilities, though, statistical inference could be a challenge. The expense of both data collection and storage prevented the acquisition and maintenance of large sample sizes, while limited access to computational power also hindered the analysis of large data sets. Undeterred, statisticians developed elegant and efficient methods that extract realistic, useful, and accurate inference out of a limited sample. These methods, as discussed in popular texts by Casella and Berger (2002) and Lehmann and Casella (2006), serve as the foundation for estimation and inference.

Key to these foundational statistics is the uniform minimum variance unbiased estimator, which we will identify as $\hat{\theta}^*$ (UMVUE). The UMVUE is an estimator that is both unbiased, $E_{\theta}\hat{\theta} - \theta = 0$, and has the smallest variance of any given estimator, $\text{Var}_{\theta}\hat{\theta}^* \leq \text{Var}_{\theta}\hat{\theta}$ (Casella and Berger, 2002). Note, though, that in the case of estimators \bar{X} and S^2 , as n gets large, $\text{Var}(\hat{\theta})$ decreases. Therefore, we can improve the variance (and thus reduce the *uncertainty*) of our estimator by increasing the size of n .

This phenomena of decreasing $\text{Var}(\hat{\theta})$ with large n leads to an interesting, yet inappropriate, conclusion for Big Data: Since uncertainty is inversely proportional to n and n is very large, estimates from Big Data always have minimal variance. This conclusion, though, requires Big Data to be housed in a single repository: one large ‘bag of data.’ On the contrary, Big Data tends to be so large that it cannot be contained in a single storage location nor processed by a single computer (Sobers, 2012). Instead, Big Data is distributed data: a lot of small ‘bags’ that together make one giant ‘bag.’ Therefore, it is incorrect and illogical to discount the existence of an

additional component of variance of an estimator attributable to large n .

Additionally, these small ‘bags’ are heterogeneous with respect to each other. Suppose that we have a set of distributed data partitioned over k smaller ‘bags.’ If the manner of partitioning is unknown, heterogeneity implies we cannot assume any or all of the k bags are the same. Instead, we only know that these bags contain the same (or very similar) collected variables. Each of the k small bags has its own structure and statistical components unique from the other bags, likely to result in a different estimate $\hat{\theta}$ for each of the k bags or partition, $\hat{\theta}_i$, $i = 1, \dots, k$. As an example, consider the demographics of the United States. Data collected from States on the West Coast are likely to vary in their statistical properties and estimates, thus differ from the same type of data collected from States on the East Coast. This heterogeneity is what makes the structure of massively distributed heterogeneous data special and peculiar. This section explores the key question: how can we learn from massively heterogeneous distributed data?

Today, “Divide and Conquer” or “Compress and Aggregate” algorithms dominate the distributed data landscape (Xi et al., 2012; McDonald et al., 2009; Liu, 1996). These algorithms are a means to get a global estimate based on a partitioning of a set of data. Suppose $Y = \{y_i : i = 1, \dots, N\}$ is a set of observed data. We identify the pooled, or centralized, estimate for this set of data as $\hat{\theta}_N$. The pooled estimate is one where we have all N observations in a single sample. This estimate is the one that minimizes the empirical risk

$$\hat{R}_N(\theta) = \frac{1}{N} \sum_{i=1}^N f(y_i, \theta), \quad (4.1.1)$$

where $f(y_i, \theta)$ denotes a loss function (Rosenblatt and Nadler, 2016). As previously discussed, though, the centralized approach is unfeasible due to the large amount of data. Furthermore, the pooled estimate may fall victim to Simpson’s paradox.

Simpson’s paradox is when the aggregate in the response differs from the patterns we observe when we analyze the individual partitions (Kievit et al., 2013; Brimacombe, 2014). We see evidence of Simpson’s paradox in modeling gene expressions, where a lack of understanding or not including certain parameters, such as clustering or multistage expression patterns, will adversely impact the accuracy and relevancy of the model no matter what statistical method is employed (Brimacombe, 2014). By blindly pooling the distributed data, we may in fact miss key insights about the heterogeneity of the data. Therefore, we need to combine parameters in a sound and careful manner that preserves the “individuality” of each partition when aggregating.

Given the concerns over pooled estimates, suppose Y can be divided into k partitions such that $Y_j = \{y_{ij} : i \in n_j\}$, where $\cup_{j=1}^k n_j = N$. Once each partition independently processes their portion of data to get their respective partition estimate $\hat{\theta}_i$, $i = 1, \dots, k$, it transmits $\hat{\theta}_i$ to an ‘oracle’ terminal for aggregation. The ‘oracle’ terminal produces the global estimate for all partitions $\hat{\theta}_i$ by collecting and aggregating the k partition estimates $\hat{\theta}_i$. Most of the research that focuses on these algorithms is communication efficiency (Rosenblatt and Nadler, 2016; Liu and Ihler, 2014; Zhang et al., 2012). An algorithm achieves optimal efficiency when it maintains a ‘one-shot’ method, where each partition sends a only a single transmission to the ‘oracle.’ Also referred to as ‘embarrassingly parallel’ learning, this technique is used by popular Big Data platforms such as MapReduce, Hadoop, and Spark (Rosenblatt and Nadler, 2016).

Current methods of conducting inference on distributed data, which concern the aggregation portion of the aforementioned algorithms, rely primarily on some form of averaging. Suppose we have k partitions of the data, the global estimate $\hat{\theta}^S$ for all partitions is found by simply averaging the mean of all partition estimates $\hat{\theta}_i$:

$$\hat{\theta}^S = \frac{1}{k} \sum_k \hat{\theta}_i, \tag{4.1.2}$$

where $\hat{\theta}_i$ is the estimate from the i^{th} partition and the S identifies it as an estimate based on the simple average. With θ as the true population parameter and $\hat{\theta}_N$ the pooled or centralized estimate that minimizes (4.1.1), Rosenblatt and Nadler (2016) showed that $\hat{\theta}^S$ is first-order equivalent to $\hat{\theta}_N$:

$$\frac{\|\hat{\theta}^S - \theta\|}{\|\hat{\theta}_N - \theta\|} = 1 + o_P(1). \quad (4.1.3)$$

The result in (4.1.3) implies that the averaged solution $\hat{\theta}^S$ converges to the true solution θ at the same rate as the pooled solution $\hat{\theta}_N$. The paper concludes that this result implies the optimality and robustness of $\hat{\theta}^S$, yet the result requires specific assumptions for both θ and $\hat{\theta}_N$. While simple averaging covers a variety of learning tasks, there are other limitations. It cannot adequately handle heterogeneity in the partitions independent of the number of observations n_i (Liu and Ihler, 2014). Also, simple averaging is not applicable either to partitions with unequal n_i or with systematic differences among partitions (non-IID) (Liu and Ihler, 2014; Rosenblatt and Nadler, 2016).

Another form of averaging is Kullback-Leibler-divergence (KL-D)-based averaging. Given an i.i.d sample $X = \{x_i : i = 1, \dots, n\}$ where X is drawn from a distribution with some unknown density $p(x|\theta)$. X is partitioned into k subsets, where each subset has its own model for its partition of X , $p(x|\hat{\theta}_k)$. The global estimate θ^{KL} is the average of the partition models with respect to a distance metric:

$$\hat{\theta}^{KL} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_i \operatorname{KL}(p(x | \hat{\theta}_i) \| p(x | \theta)) \quad (4.1.4)$$

where $\operatorname{KL}(p(x | \hat{\theta}_i) \| p(x | \theta)) = \mathbb{E}_{p(x|\hat{\theta}_i)} \left[\log \frac{p(x|\hat{\theta}_k)}{p(x|\theta)} \right]$. This method averages $\hat{\theta}$ based on a distance metric and not the parameters themselves (Liu and Ihler, 2014).

While Liu and Ihler (2014) shows $\hat{\theta}^{KL}$ achieves minimal bias and minimal mean squared error (MSE) for global estimates of $\hat{\theta}$, it does so under some limitations. First,

this method only applies when the density $p(x | \theta)$ comes from a curved exponential family distribution. These distribution are detailed in Efron (1975). Second, KL-D is still more robust than linear averaging when the model is misspecified; unfortunately, it cannot correct the misspecification.

Finally, Minsker and Strawn (2017) addresses the condition for large k . When the number of partitions is large, they identify the requirement for a more robust ‘combiner’ than just averaging to effectively manage any anomalous partition estimates $\{\hat{\theta}_i\}$. Minsker and Strawn (2017) proposes the ‘median of means’ as a more robust approach:

$$\hat{\theta}^{\text{med}} = \text{med}(\bar{\theta}_1, \bar{\theta}_2, \dots, \bar{\theta}_k),$$

where $\bar{\theta}_i$ represents the local mean estimates for each of the k partitions. While large values of k do not significantly impact the algorithm (Minsker and Strawn, 2017), it does require two key assumptions about the local estimators, $\bar{\theta}_i$, $i = 1, \dots, k$. First, all $\bar{\theta}_i$ are IID from some common distribution P_k . Second, P_k is approximately symmetric. Thus, the distance between the median of (P_k) and the mean (P_k) is small.

4.1.1 V-Data

Of the three methods detailed here, none of them effectively handles unconstrained heterogeneity in distributed data. Each method requires one or more assumptions on the data, yet the vast size and heterogeneity of distributed data prevents such restrictions. Consider the example in Figure 4.1. Here, we have $N = 5000$ total observations equally divided into $k = 50$ partitions. Each partition includes data that includes observations that behave in a similar manner: either the partition’s data trends negatively (the left side of the figure) or positively (the right side). Although we can readily observe this heterogeneity in the data set, we understand that detecting

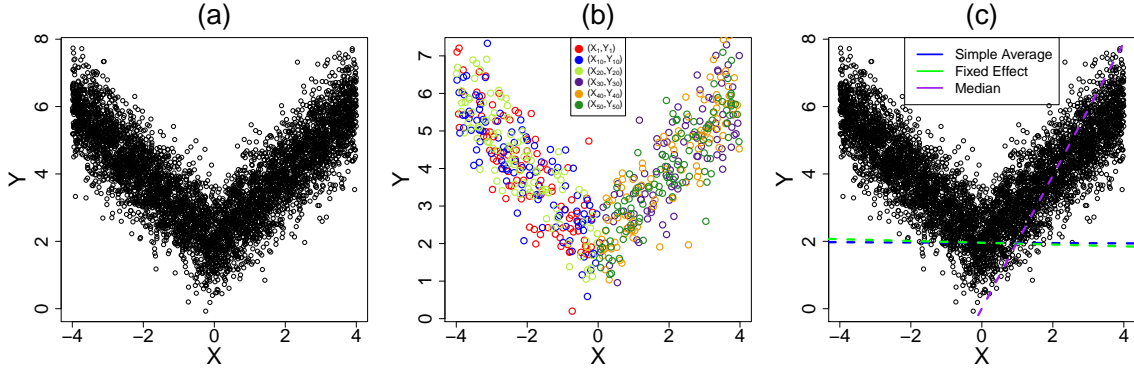


Figure 4.1: An illustration of an extreme example in massively heterogeneous distributed data. Panel (a) shows the complete set of 5000 observations, while panel (b) gives an example of the 50 observations in a specific partition k . Panel (c) illustrates how simple averaging, fixed effect modeling, and the median of modes estimates the relationship based on partition estimates.

this heterogeneity in an actual case of distributed data is significantly challenging.

Given $\mathbb{E}[Y | X = x] - \bar{Y} = \beta X$, each partition estimates own significant $\hat{\beta}$ that results in a value close to either positive 1 or negative 1. When we apply some of the aforementioned methods of combining estimates, though, we see a different story. In Figure 4.1(c), we see the results of simple averaging, fixed effect modeling, and the median of means. In the first two cases, the global estimates for $\hat{\beta}$ based on the methods we previously discussed says X is not predictive of Y . As for the median of means, we see how positive X is predictive of Y , but the model fails to capture the true shape of the data when X is negative. The limitation of the aforementioned methods is that they seek a single $\hat{\beta}$ to describe the entire set of data. In this example, though, a single $\hat{\beta}$ is inappropriate.

We can use macroinference of the DS prior as an adaptive inference method to determine a grand, ‘global’ estimate for a population parameter θ with little or no restrictions on the types of data or its structure. In this example of distributed data, the gEB and the DS prior will allow us to capture the true functional shape of the distribution of the β parameters. Furthermore, the generalized empirical Bayes approach provides an embarrassingly parallel method to combine k estimates, enables

improvements to local estimates through elastic shrinkage, and allows us to quantify the heterogeneity. Most importantly, the non-parametric behavior of the DS prior will grant the flexibility to account for the extreme shape in the data. Figure 4.2 shows the results when we apply generalized empirical Bayes to the ‘V-Data’ scenario.

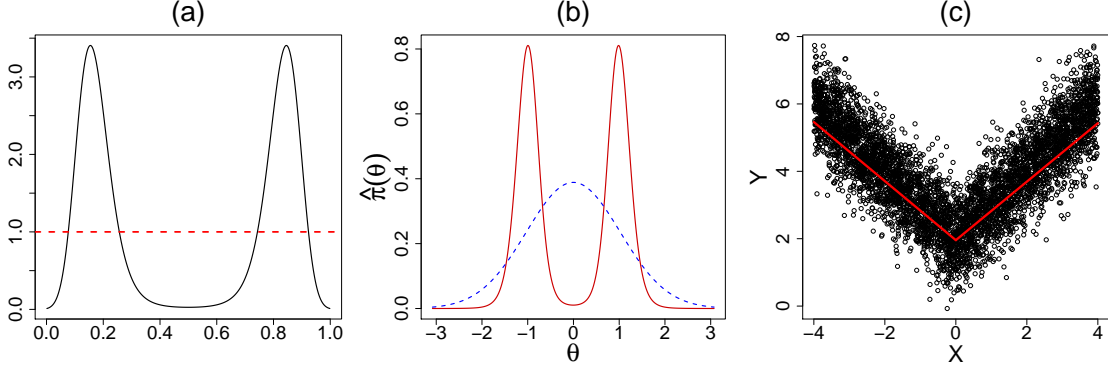


Figure 4.2: Panel (a) shows the U-function for the ‘V-Data’ problem, which tells us that there are two distinct modes. Panel (b) provides a comparison of the DS(G,m) prior versus the parametric prior for β . Panel (c) shows how the modes of (b) estimate the true nature of the data.

The uncertainty function in Panel (a) of Figure 4.2 indicates that the assumption of normality for β is inadequate and needs significant repair. The large bimodality in the uncertainty function indicates that we can expect a similar change to the functional shape of our assumption, which is clear in Panel (b) of Figure 4.2. After conducting macroinference, we find that the two modes are $\beta = -0.99 \pm 0.108$ for values of X less than zero and $\beta = 0.99 \pm 0.108$ for values of X greater than zero. When we incorporate these results into our final model for the V-data, we get the model shown in Figure 4.2(c):

$$\mathbb{E}[Y | X = x] = \begin{cases} -0.99X + 1.96, & X \leq 0, \\ 0.99X + 1.96, & X > 0. \end{cases} \quad (4.1.5)$$

While the example here is extreme in shape, gEB and the DS prior succeeded in providing a more representative model. The example enabled us to show another

application of gEB: distributed regression modeling. Unlike standard regression that pools the data to generate an answer, distributed regression modeling combines parameter estimates from each partition for a global estimate. As shown, other methods of distributed regression modeling will produce poor models if the method of combining does not incorporate mechanisms to identify the heterogeneity in the data and adjust accordingly. Generalized empirical Bayes is a framework well suited to distributed regression modeling because it is designed to capture the uniqueness of each partition and utilize its heterogeneity to generate an improved and insightful model. The next example will demonstrate such a result.

4.1.2 Cheese Data

In the same spirit as the V data, we continue our applications of gEB on distributed data sets with non-simulated data. The ‘cheese’ dataset, from the `bayesm` package in R, includes 5,555 observations of weekly sales volume for a package of Borden sliced cheese for $k = 88$ stores. Additionally, it has two independent variables: the measure of promotional activity at the store and the price of a package. Our goal is to compare traditional regression modeling to distributed regression modeling.

In traditional regression modeling, we ignore any heterogeneity between the $k = 88$ stores and pool the data to develop a global model. On the contrary, distributed regression modeling with gEB respects and smartly incorporates the heterogeneity between the stores to arrive at a more descriptive and accurate global model. Is there heterogeneity in the cheese data? Figure 4.3 shows scatter plots of the relationship between the promotional activity and sales volume for three of the $k = 88$ stores. Each plot shows a significantly different estimate for the target relationship; the data is heterogeneous and requires more than a pooled regression approach.

We proceed with both a pooled regression and distributed regression. For this analysis assume a linear model, with the log of sales volume y_i as our dependent

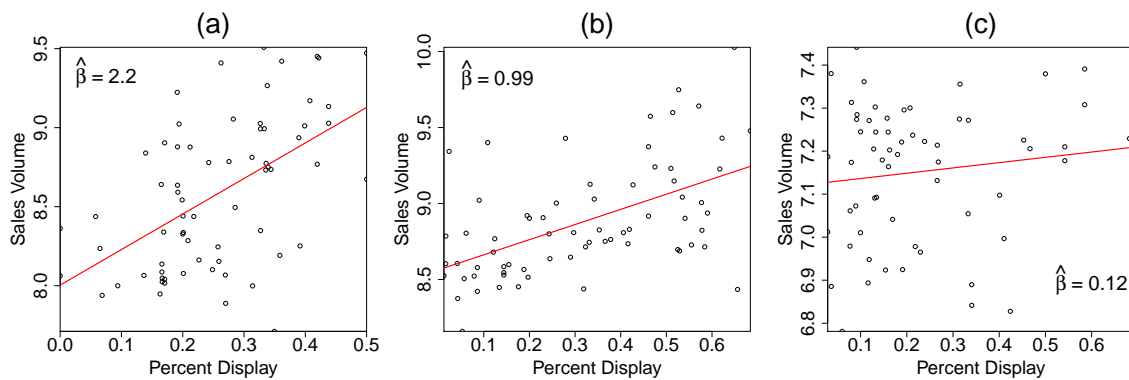


Figure 4.3: Scatter plots, regression lines, and $\hat{\beta}$ for three of the $k = 88$ stores in the cheese data. Each panel shows a different relationship between the promotional activity (percent display) and the sales volume.

variable and the log of price x_1 and promotional activity x_2 as independent variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad (4.1.6)$$

where $\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$. To better compare with gEB, we will use Bayesian regression methods which requires a prior distribution on the parameters β_0 , β_1 , and β_2 . We will use a $\mathcal{N}(0, 10)$ prior for each β ; this diffuse, or variance-inflated, prior provides some structure and prevents our distributional assumption from inadvertently constraining the model. Finally, we assume that there is no correlation between our independent variables. This assumption is key in regression, as it allows for no multicollinearity in the data. It is even more important in distributed regression; in order to combine the partition parameter estimates using gEB, those estimates must be independent.

For Bayesian regression, we will use posterior sampling through Markov Chain Monte Carlo simulation to determine the parameter's distribution, estimate, and standard error. These functions are captured in the R package `bayesm`. For the distributed regression, we will follow a similar process as used in the V-data example. First, we treat each store as a partition with its own set of observations. Next we build a regression model (4.1.6) for each store using its own observations, resulting in $k = 88$ models. For each of the three parameters, we will combine them using gEB

for a global model. The results for both the Bayesian pooled regression and the gEB distributed regression are shown in Table 4.1.

Table 4.1: Comparison of parameters for the cheese data derived from a pooled Bayesian regression and a distributed embarrassingly parallel framework.

Parameter	Method	Mean	St Dev
Intercept	Pooled Regression	9.37	0.063
	Distributed Regression	10.3	0.109
log(Price)	Pooled Regression	-1.26	0.058
	Distributed Regression	-2.1	0.074
Display		0.51	0.064
		0.44	0.194
	Distributed Regression	0.995	0.092
		1.489	0.133

Initially, we see that both methods preserved the relationship between the individual parameters and the log of sales volume. The parameter $\log(\text{Price})$ is negatively correlated with the log of sales volume, which indicates that higher prices results in fewer sales. The display parameter has a positive relationship with the dependent variable means that more space dedicated to promotion display increases the amount of sales. When we look at the values for the parameters, we notice a difference between the pooled and distributed regression. In terms of $\log(\text{Price})$, the distributed regression shows a much larger penalty for a price increase than the pooled; the majority of partition parameters ranged between -5 and 0, but there were two cases where stores had a parameter value smaller than -8 and two cases where stores had values greater than 2. Since each partition has between 60 and 70 observations, the pooled regression model would treat each observation equally while the distributed regression concentrates on the distribution of the parameter among the stores. The most significant finding, though, lies with the display parameter.

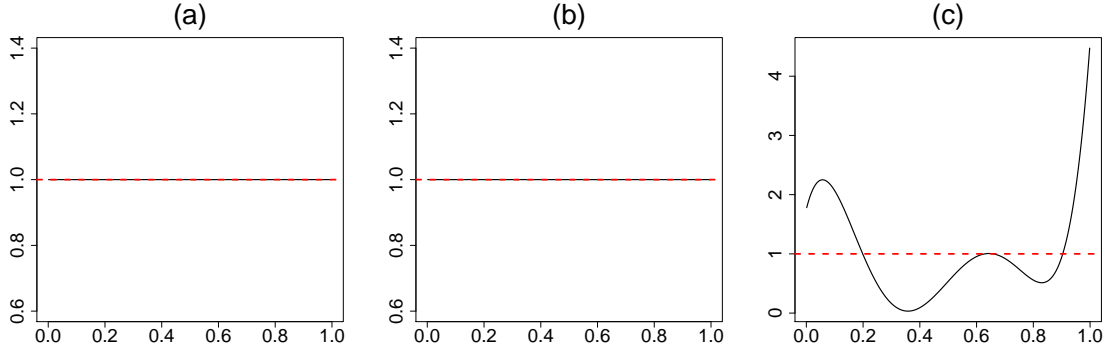


Figure 4.4: The U-functions for the parameters of the cheese data, when analyzed under a distributed structure. Panel (a) is the intercept, (b) is the log(price), and (c) is the display.

In Table 4.1, we see the pooled regression model produces a positive value for the display parameter β_2 . In contrast, we have three *distinct* values resulting for the distributed regression model. To understand why we have such a result, one must look at the uncertainty functions in Figure 4.4. The flat lines for Panels (a) and (b) show $d = 1$ for both the intercept β_0 and the log(Price) β_1 . Panel (c) is the uncertainty function for the display parameter; the tri-modality of the uncertainty function suggests that our assumption of normality for β_2 is not adequate and requires repair. The resulting equation for the DS prior of the display parameter, with $g(\theta) \sim \mathcal{N}(0.873, 0.139)$ is

$$\hat{\pi}(\theta) = g(\theta) [1 + 0.651T_2(\theta; G) + 0.234T_4(\theta; G) + 0.423T_5(\theta; G)] \quad (4.1.7)$$

Figure 4.5 shows a comparison of histograms resulting from posterior sampling of parameters in the pooled regression to the distributed regression parameter distributions. We have already acknowledged the difference in means between the intercept and log(price), which are indicated in Panels (a) versus (d) and (b) versus (e). Our focus is in the comparison between Panels (c) and (f).

In Panel (c), we see the results of a centralized pooling of the partition data. The

result from the distributed regression model in Panel (f), though, shows that the centralized estimate is plagued by Simpson’s paradox. The shape of the distribution indicates that there exist three different groups of stores with respect to how they display the cheese product. Of the three modes, $\beta_2 = 0.444$ is the maximum mode while $\beta_2 = 0.995$ is the most precise mode (standard error of 0.092). Through the application of microinference to the partition estimates of β_2 , we can improve the parameter estimates and subsequently apply k-means clustering to identify which stores fall into particular groups: 21 stores align with the $\beta_2 = 1.5$, 22 stores align with $\beta_2 = 0.44$, and 43 correspond to $\beta_2 = 0.995$. The detailed feedback from distributed regression enables company leadership and store management to better identify areas of interest. For those stores who get more out of their displays ($\beta_2 = 1.5$), the company would want to identify how they achieve such high rates and share those practices with other stores. On the other hand, leadership can pinpoint those stores who are not maximizing their marketing displays and investigate.

4.1.3 Key Observations

In this section, we showed how generalized empirical Bayes provides a unique and acceptable alternative when seeking inference over distributed data. Unlike other embarrassingly parallel schemes, gEB and distributed regression modeling embraces the heterogeneity among the partitions. Through this embrace, we can efficiently produce models that better describe the given situation (V-data) and give decision makers better insight (cheese data). Not only does this method of distributed modeling facilitate efficient computing, but it also prevents the disclosure of proprietary or personal information. In both examples, we produced a model based on parameter estimates from the partitions instead of utilizing the partition’s full data. The combination of gEB and distributed modeling permits a *privacy-preserving* form of data analysis that can facilitate the sharing of data without compromising Personally Iden-

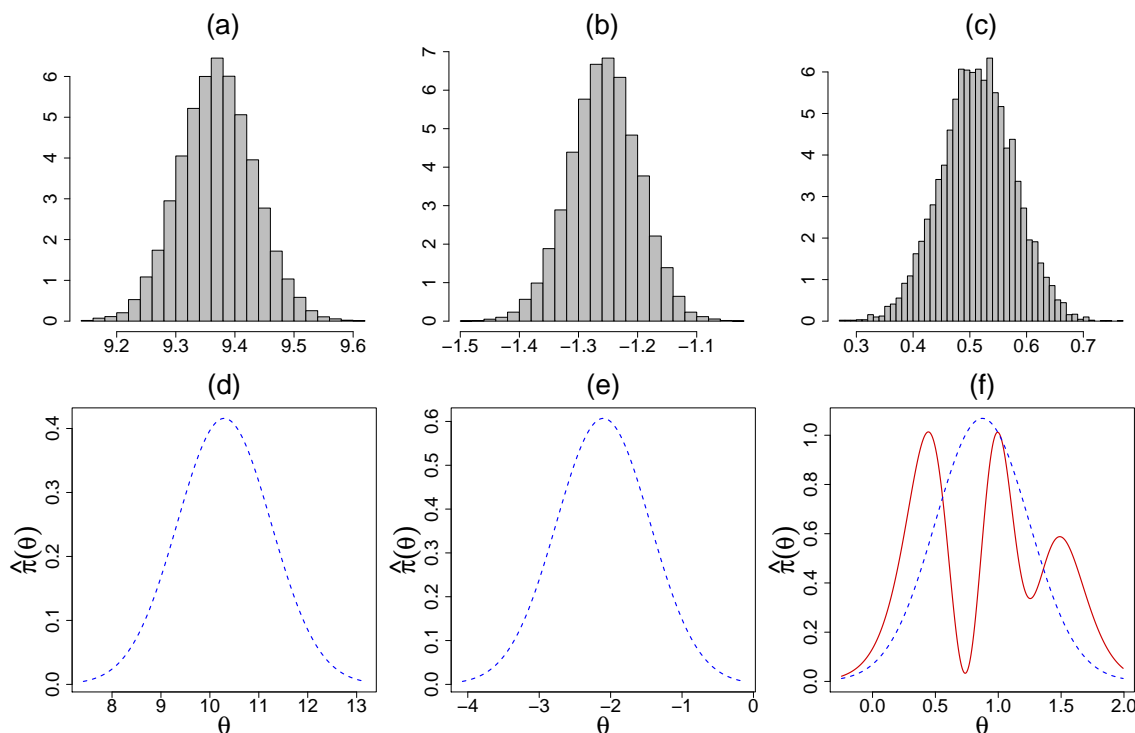


Figure 4.5: A comparison of the distributions of the regression parameters between pooled data using Bayesian regression (panels (a) through (c)) and distributed data using DS prior (panels (d) through (e)). The comparison of panel (c) and panel (f) indicates the presence of Simpson’s paradox in the pooled analysis.

tifying Information (PII) or devaluing data collected by corporations or businesses.

4.2 The Missing Species problem

In Chapter 2.3.4, we introduced Corbett’s Butterfly data (Fisher et al., 1943) and demonstrated how the DS prior provides a smoother microinference for $\mathbb{E}(\theta_i \mid y_i)$ when compared to other Bayes estimates. Corbett’s critical question, though, dealt with the number of new species he could potentially capture given additional time (Fisher et al., 1943; Efron and Hastie, 2016). In other words, consider the Malaysian butterfly population is composed of an unknown number of species S . While the number of butterflies that belong to each species may be of interest, the focus of Corbett’s question is estimating the true size of S .

We assume that the population is sampled based on some measure of time t . In terms of Corbet’s butterfly data, t represents the two-year interval he spent trapping butterflies. Table 4.2 provides the details on the counts for the species seen y times during the two years, where $y \in \{1, 2, \dots, 23, 24\}$.

Table 4.2: The butterfly data, where the count of species seen y times during Corbet’s two-years of trapping.

y	1	2	3	4	5	6	7	8	9	10	11	12
Counts	118	74	44	24	29	22	20	19	20	15	12	14
y	13	14	15	16	17	18	19	20	21	22	23	24
Counts	6	12	6	9	9	6	10	10	11	5	3	3

Let S be the true number of species, both observed and unobserved, while N is the number of only the observed species such that $N \subset S$. As with Fisher et al. (1943) and Efron (2017), we assume that the number of times a species is trapped in interval t follows a Poisson distribution:

$$y_i | \theta_i \sim \text{Poisson}(\theta_i), \quad i = 1 \dots N \tag{4.2.1}$$

$$\theta_i \stackrel{\text{ind}}{\sim} g(\theta)$$

We frame Corbett’s question in terms of the model presented in (4.2.1): given an additional year spent trapping butterflies ($t = 0.5$), how many *new species* can Corbett expect to capture? Our model assumes that each species has its own θ_i , where θ_i is the expected number of species i captured over a single time unit t (in the case of Corbett, that time unit is 2 years).

To address Corbett’s question, we formulate the probability that a species *is not* caught during the initial period, $\mathcal{P}(Y_i = 0)$, but *is* caught in the next t periods, $\mathcal{P}(Y_i > 0)$. Recognizing these probabilities are independent events from a Poisson

process, our final probability is

$$e^{-\theta_i}(1 - e^{-\theta_i t}) \tag{4.2.2}$$

Let $u(t)$ be the expected number of new species seen in the next t time periods.

$$\begin{aligned} u(t) &= \sum_{i=1}^S e^{-\theta_i}(1 - e^{-\theta_i t}) \\ &= S \int_0^{\infty} e^{-\theta}(1 - e^{-\theta t})g(\theta) d\theta \end{aligned} \tag{4.2.3}$$

Equation (4.2.3) provides a time based solution to estimating the number of new species Corbett can expect to capture in t additional periods. The importance of this question spans more than just the Malaysian butterfly population. We can easily extend Corbett’s question to the estimation of any species in populations of plant and animal life to gain insight into extinction rates and manage biodiversity (Bunge and Fitzpatrick, 1993; Chao, 2005). We can also use a similar model to estimate an author’s complete vocabulary of distinct words (Efron and Thisted, 1976; Thisted and Efron, 1987; Orlitsky et al., 2016). As an example, consider the famous playwright and poet William Shakespeare. We have volumes and volumes of his works that document his confirmed lexicon; yet, does that documentation cover all of the words he truly knew (Efron and Thisted, 1976; Efron and Hastie, 2016)? Bunge and Fitzpatrick (1993) provide even more applications, spanning from estimating the number of dies used to mint coins to unique records in a filing system. Although commonly referred to as the ‘missing species problem’ or ‘species estimation,’ the vast scope of its applications emphasizes its importance. Essentially, the missing species problem provides a method of estimating the number of true partitions S in a given population (Bunge and Fitzpatrick, 1993).

The goal of this section is to demonstrate how we can use generalized empirical

Bayes (gEB) and the $DS(G, m)$ prior to address the ‘missing species problem.’ We will pursue two distinct approaches: g -modeling and f -modeling. The g -modeling approach requires us to estimate the prior distribution $g(\theta)$ in our model (4.2.1). The second approach, f -modeling, relies on the estimation of the marginal density $f_G(y)$. We begin by discussing the use of DS priors in the g -modeling approach and its application to Corbett’s problem. Afterward, we will illustrate the f -modeling approach where we will the DS distribution introduced in Chapter 3. In addition to Corbett’s butterfly data, we also apply our approach to estimating Shakespeare’s total canon and predicting the number of words in Hamlet.

4.2.1 g -Modeling

The most popular g -modeling approach was that by R.A. Fisher (Fisher et al., 1943), who took a parametric empirical Bayes approach to answer Corbett’s question. Another method, empirical Bayesian deconvolution, is detailed in Efron (2016) and is a complete nonparametric approach to estimating $g(\theta)$. Since our methodology relies on the identification of a parametric g , we will focus on Fisher’s approach.

Fisher’s PEB approach

Fisher will begin with 4.2.3, but with a slight modification. From the previous section, we know $e_x = \mathbb{E}(y_x) = S \int_0^\infty \frac{e^{-\theta} \theta^x}{x!} g(\theta) d\theta$. Let $x = 1$, then

$$e_1 = S \int_0^\infty e^{-\theta} \theta g(\theta) d\theta. \quad (4.2.4)$$

Solving 4.2.4 for S and substituting it into equation 4.2.3, we get the starting point for Fisher’s derivation:

$$u(t) = \frac{e_1 \int_0^\infty e^{-\theta} (1 - e^{-\theta t}) g(\theta) d\theta}{\int_0^\infty \theta e^{-\theta} g(\theta) d\theta} \quad (4.2.5)$$

Fisher, though, let g be the gamma distribution such that $g(\theta) = C_{\alpha,\beta}\theta^{\alpha-1}e^{-\theta/\beta}$. He then substituted it into 4.2.5:

$$\begin{aligned}
u_G(t) &= \frac{e_1 \int_0^\infty e^{-\theta}(1 - e^{-\theta t})g(\theta) \, d\theta}{\int_0^\infty \theta e^{-\theta}g(\theta) \, d\theta} \\
&= \frac{e_1 \left[\int_0^\infty \theta^{\alpha-1} e^{-\theta(\frac{1+\beta}{\beta})} \, d\theta - \int_0^\infty \theta^{\alpha-1} e^{-\theta(\frac{1+\beta+t\beta}{\beta})} \, d\theta \right]}{\int_0^\infty \theta^{(\alpha+1)-1} e^{-\theta(\frac{1+\beta}{\beta})} \, d\theta} \\
&= \frac{e_1 \left[\Gamma(\alpha) \left(\frac{\beta}{1+\beta} \right)^\alpha - \Gamma(\alpha) \left(\frac{\beta}{1+\beta+t\beta} \right)^\alpha \right]}{\alpha \Gamma(\alpha) \left(\frac{\beta}{1+\beta} \right)^{\alpha+1}}
\end{aligned}$$

Let $\gamma = \frac{\beta}{1+\beta}$:

$$\begin{aligned}
u_G(t) &= -\frac{e_1}{\alpha\gamma} \left[\left(\frac{\beta+1}{\beta+t\beta+1} \right)^\alpha - 1 \right] \\
&= -\frac{e_1}{\alpha\gamma} \left[\frac{1}{\left(\frac{\beta+1}{\beta+1} + \frac{t\beta}{\beta+1} \right)^\alpha} - 1 \right] \\
&= -\frac{e_1}{\alpha\gamma} [(1 + \gamma t)^{-\alpha} - 1] \tag{4.2.6}
\end{aligned}$$

Allow $\hat{e}_1 = y_1 = 118$, the maximum likelihood estimation results in $\hat{\alpha} = 0.104$ and $\hat{\gamma} = 0.9886$ (Efron and Hastie, 2016). Substituting those values, along with $t = 0.5$, into (4.2.6) gives $\mathbb{E}(0.5) = 46.95$. Efron and Hastie (2016) comments that Fisher's approximation also provides reasonable approximations for $t > 1$. Furthermore, Efron and Thisted (1976) show that (4.2.6), with $\hat{\alpha} = -0.395$, $\hat{\gamma} = 0.991$, and $e_1 = 14376$, gives a sensible estimate for the number of distinct words in Shakespeare's lexicon: $\mathbb{E}_{\text{SHK}}(1) = 11483$.

Efron's Revision to Fisher's Equation

In the butterfly problem, as with all forms of the missing species problem, we know there exists a complete population of a S species each with its own Poisson parameter θ_i . In Corbett's two years, though, he found only $N = 501$ of the S total species. According to Efron (2017), Corbett only found butterflies where $y_i \sim \text{Poisson}(\theta_i)$ was greater than zero (we will ignore any truncation for $y_i > 24$). As such, let $g^+(\theta)$ be the prior density that applies to all species S . Since $\mathcal{P}(y_i > 0 \mid \theta_i) = 1 - e^{-\theta_i}$, then

$$g(\theta) = c(1 - e^{-\theta})g^+(\theta) \quad (4.2.7)$$

where c is a normalizing constant. We can solve for c by looking at the expected number of species that Corbett observed:

$$\begin{aligned} \mathbb{E}[N] &= S \int g^+(\theta)(1 - e^{-\theta}) \, d\theta \\ &= S \int \frac{1(1 - e^{-\theta})}{c(1 - e^{-\theta})} g(\theta) \, d\theta \\ &= \frac{S}{c} \approx N \rightarrow c = \frac{S}{N} \end{aligned} \quad (4.2.8)$$

In Equation 4.2.5, we assume that $g(\theta) = g^+(\theta)$. To update 4.2.5 as in Efron (2017), we substitute 4.2.7 and 4.2.8 and arrive at the following result:

$$u_{\text{EF}}[t] = N \int e^{-\theta} \frac{(1 - e^{-\theta t})}{(1 - e^{-\theta})} g(\theta) \, d\theta \quad (4.2.9)$$

gEB and g -Modeling

Now, we will use generalized empirical Bayes and the DS prior to estimate the number of new species captured given an extra year. Suppose we begin with Equation 4.2.9

and, instead of $g(\theta)$, we choose to use the DS prior $\pi(\theta)$:

$$\begin{aligned} u_{\text{DS}}[t] &= N \int e^{-\theta} \frac{(1 - e^{-\theta t})}{(1 - e^{-\theta})} \pi(\theta) \, d\theta \\ &= N \int e^{-\theta} \frac{(1 - e^{-\theta t})}{(1 - e^{-\theta})} g(\theta) d[G(\theta)] \, d\theta \end{aligned}$$

Let $G(\theta) = u$ where $0 \leq u \leq 1$:

$$u_{\text{DS}}[t] = N \int e^{-[G^{-1}(\theta)]} \frac{(1 - e^{-[G^{-1}(\theta)]t})}{(1 - e^{-[G^{-1}(\theta)]})} d[u] \, du \quad (4.2.10)$$

The result in (4.2.10) gives us a remarkably tractable integral in terms of the total number of observed species N , the quantile function $G^{-1}(\theta)$ and the uncertainty function $d[u]$. We only need to estimate $d[u]$ to solve the integral, which is easily done with the Type-II Method of Moments algorithm detailed in Chapter 2.

Prior to applying the algorithm, though, we must reconsider the selected model in (4.2.1). In the case of the butterfly data, $y \in \{1, 2, 3, \dots, 24\}$; the exclusion of zero and $y > 24$ illustrate that we are dealing with double truncated data. Since our algorithm requires the expectation with respect to the posterior, we consider the impact of truncation on the posterior. Given

$$\pi_G(\theta_i | y_i) = \frac{f(y | \theta_i)g(\theta_i)}{f_G(y)}$$

we know that, under truncation, the likelihood $f(y | \theta_i)$ is a double truncated Poisson distribution and the marginal $f_G(y)$ is a double truncated negative binomial. As such, we account for truncation by dividing out the probability mass over the entire truncation range (in the case of the butterfly, our truncation range would go from 1 to 24). The resulting value, the sum of all the probability mass over the given truncated range, will either be a constant or a function of θ . For our marginal

distribution, we have a constant value C_{TR} . The probability mass for the likelihood, though, is dependent upon θ_i . Therefore, we denote it at $P_{\text{TR}}(\theta_i)$. Our final posterior distribution given truncation is $\frac{C_{\text{TR}}}{P_{\text{TR}}(\theta_i)}\pi_G(\theta_i)$. Let

$$h^*(\theta_i) = \frac{C_{\text{TR}}}{P_{\text{TR}}(\theta_i)} \text{Leg}_j(G(\theta_i)). \quad (4.2.11)$$

With our updated $h^*(\theta_i)$, we can update \mathbb{E}_G in (2.2.3) as follows:

$$\begin{aligned} \mathbb{E}_G [h^*(\theta_i)] &= \mathbb{E}_G \left[\frac{C_{\text{TR}}}{P_{\text{TR}}(\theta_i)} \text{Leg}_j(G(\theta_i)) \right] \\ &= \int \frac{C_{\text{TR}}}{P_{\text{TR}}(\theta_i)} \text{Leg}_j(G(\theta_i)) \pi_G(\theta_i) d\theta_i \end{aligned}$$

As before, we let $G(\theta) = u$ where $0 \leq u \leq 1$:

$$\mathbb{E}_G [h^*(\theta_i)] = \int \frac{C_{\text{TR}}}{P_{\text{TR}}(G^{-1}(\theta))} \text{Leg}_j(u) \frac{\pi_G(G^{-1}(\theta))}{g(G^{-1}(\theta))} du$$

Now, we can easily update our algorithm with h^* and adapt it to the missing species problem. As with previous examples in this dissertation, we simply execute the algorithm to determine any appropriate $\text{LP}[\theta; G, \Pi]$ parameters for problem. Prior to executing the algorithm, we must first determine an appropriate g .

Determining g

Since the likelihood of our model follows a Poisson distribution, a logical selection is the conjugate prior distribution $g(\theta) \sim \text{Gamma}(\alpha, \beta)$. From Table 2.1, we know the marginal distribution $f_G(y)$ is the negative binomial distribution. It follows that we can use MLE to determine the appropriate parameters, but we must show some caution. As mentioned in the previous section, the model deals with a truncated distribution. Consider Figure 4.6, which shows the maximum likelihood estimates for g with no truncation, zero truncated, and double truncated. The estimated prior with-

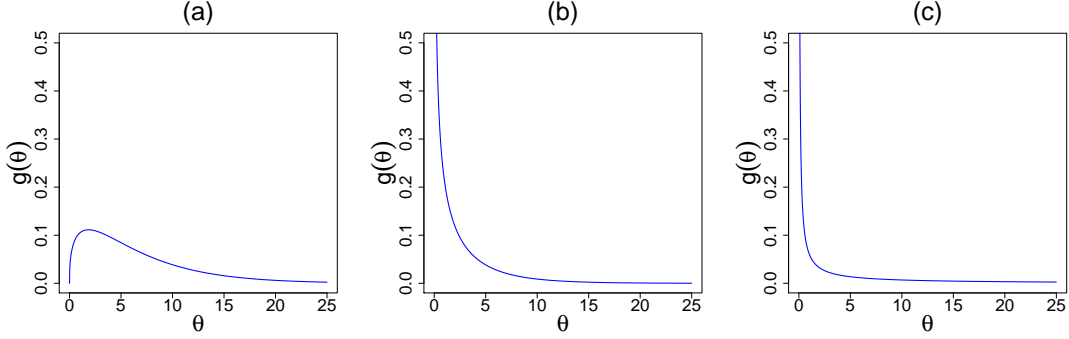


Figure 4.6: Three different candidates for $g(\theta)$: (a) no truncation; (b) zero truncation; and (c) double truncation.

out considering truncation (Figure 4.6(a)) portrays a completely different scenario than the others. Although the $DS(G, m)$ class of priors will repair the given g should prior-data conflict exist, it would require much larger values of m which could result in over-fitting. Instead, we want to smartly select g so that we provide the algorithm a starting point that considers the nuances of the given data. Blindly applying MLE, while an option, does not properly account for the truncation and could arrive at a starting g that may prove troublesome.

Given these considerations, we can alternatively treat the data as a multinomial with y_0 categories, where $y_0 \in \{1, 2, \dots, N\}$ and the probabilities p_y for a category y is proportional to

$$p_y \propto \frac{\Gamma(y + \alpha)}{y! \Gamma(1 + \alpha)} \gamma^{y-1}. \quad (4.2.12)$$

This construct follows from the perspective detailed earlier that involves determining the true size of S . We consider the observed number of species N as the maximum available categories or bins available. Now, p_y represents the probability that we observed the given category; we want the parameters α and $\gamma = \frac{\beta}{1+\beta}$ that maximize p_y given the observed data. Under this assumption, we can now find the optimal parameters for $g(\theta)$ by only considering a subset of y 's such that $y_0 < N$. We will use the work of Orlitsky et al. (2016), as detailed in the f -modeling section, to assist in guiding our selection of y_0 .

Analysis of Butterfly data

While the nature of the model in (4.2.1) restricts us to the family of gamma distributions, we have flexibility in determining the parameters. For the butterfly data, we incorporate the previous points into a more scientific approach to determining a suitable $g(\theta)$ prior to the application of MLE for parameter estimation. We want g to reflect the commonality of the abundant species and account for those species that are more uncommon and less likely to be found. These particular conditions lead us to fix $\alpha = 1$ and use an exponential distribution with parameter β . Since the exponential distribution is a form of gamma distribution, we do not require any further modification to the algorithm.

Now that we have determined the initial structure of $g(\theta)$, we want to find the estimate for β . We will use the approach from Efron and Thisted (1976) summarized in (4.2.12). To determine our truncation point y_0 , we use the recommended truncation for the Efron-Thisted estimate outlined in Orlitsky et al. (2016) that sets $y_0 = 6$. This method allows us to use the data on the more uncommon species to estimate our distribution's parameter, which should assist in providing more accurate estimates for prediction. After applying the algorithm, we have the $DS(G, m = 8)$ prior where g follows an exponential distribution ($\hat{\beta} = 2.16$), in its \mathcal{L}^2 representation:

$$\begin{aligned} \pi(\theta) = g(\theta) & \left[0.62T_1(\theta; G) + 0.49T_2(\theta; G) + 0.49T_3(\theta; G) + 0.52T_4(\theta; G) \right. \\ & \left. + 0.50T_5(\theta; G) + 0.43T_6(\theta; G) + 0.35T_7(\theta; G) + 0.25T_8(\theta; G) \right] \end{aligned} \quad (4.2.13)$$

Next, we use this model to conduct both microinference and prediction. Figure 4.7 shows the results. In panel (a), we have estimates for the number of butterflies caught in the following year for $y \in \{1, 2, \dots, 23, 24\}$. Note how the elastic-Bayes estimate for the DS prior fits between the parametric exponential distribution and the nonparametric deconvolution estimate. Initially, the DS estimates follows the

nonparametric deconvolve for the less-common species. As y increases, the resulting expectation becomes closer to the parametric estimate for those species that are more common. When compared to our previous result in Figure 2.4, which used the gamma prior with MLE parameters from Fisher’s derivation in (4.2.6), we see that the DS Prior with $g(\theta)$ as an exponential gives a more desirable result that borrows from both parametric and non-parametric.

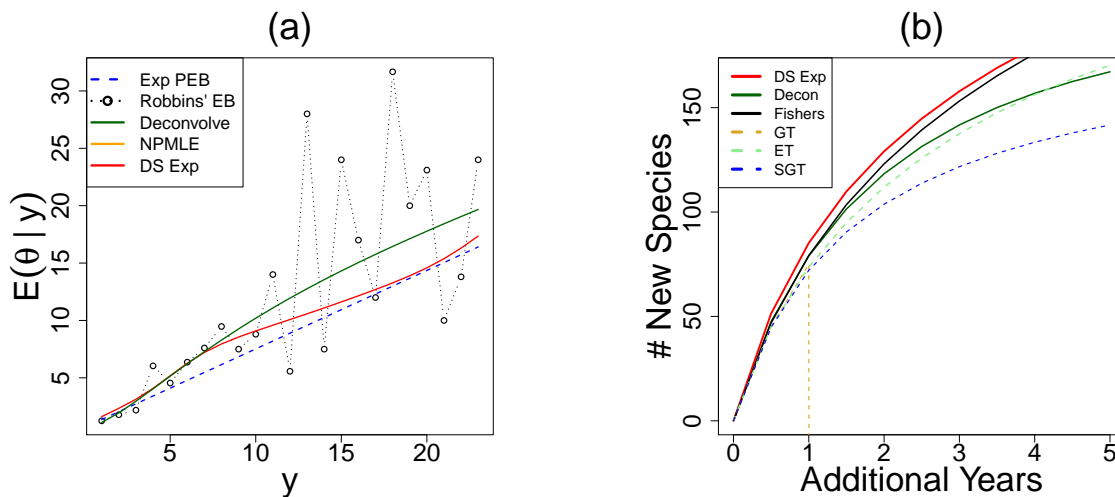


Figure 4.7: Panel (a) shows estimates for the number of butterfly species caught in the following year $\hat{E}(\theta | y)$ by the Exponential PEB, Robbins’ formula, Bayesian deconvolution, NPMLE, and our Elastic-Bayes estimate using the Exponential starting priors. Panel (b) shows the prediction of new number of species for additional years.

Discussion of Results

We began with the g -modeling approach because its foundation is the prior distribution and easily aligns with the DS prior construct. With the butterfly data, we see that the DS prior approach provides an alternative that falls between Fisher’s completely parametric approach and Efron’s nonparametric Bayesian deconvolution approach. While these results are encouraging, we must consider the difficulty in selecting an appropriate g to begin the analysis. In many situations, the MLE estimates for the parameters of g are more than appropriate. In the case of truncation, though,

we find that the MLE may not provide the best starting point. As we discussed above, we took significant input from the ‘subject-matter expert’ in order to arrive at an appropriate result. While relying on the ‘human-in-the-loop’ is not detrimental to our findings, we want to explore methods that require less effort from the subject-matter expert. One alternative, and an easier approach, is to bypass g entirely and use the observed data y to construct the model. We now transition into extending the DS prior concept to support an f -modeling approach.

4.2.2 f -Modeling

The f -modeling approach relies on the estimation of the marginal density $f_G(y)$. Some modern techniques include Lindsay’s method (Efron, 2012, 2017), a linear program approach to estimating the densities for the unobserved portion (Valiant and Valiant, 2013), and the use of Chebyshev polynomials to estimate the support size of distribution (Wu and Yang, 2015; Orlitsky et al., 2016). One of the earliest methods of f -modeling is the Good and Toulmin estimator (Good and Toulmin, 1956; Efron and Thisted, 1976; Orlitsky et al., 2016). This estimator utilizes one of the earliest examples of empirical Bayes: Robbins’ formula. Introduced in Robbins (1956), Robbins’ formula sought to approximate the expectation of a Poisson random variable given an observation y . According to Bayes rule, we know that the posterior takes the form

$$f(\theta | y) = \frac{f(y | \theta)g(\theta)}{f(y)} = \frac{\frac{e^{-\theta}\theta^y}{y!}g(\theta)}{\int_0^\infty \frac{e^{-\theta}\theta^y}{y!}g(\theta) d\theta} \quad (4.2.14)$$

We take the expectation of the posterior and find

$$\begin{aligned}
\mathbb{E}(\theta \mid Y = y) &= \frac{\int_0^\infty \theta \frac{e^{-\theta} \theta^y}{y!} g(\theta) \, d\theta}{f(y)} \\
&= \frac{(y+1) \int_0^\infty \frac{e^{-\theta} \theta^{(y+1)}}{(y+1)!} g(\theta) \, d\theta}{f(y)} \\
&= \frac{(y+1)f(y+1)}{f(y)} \tag{4.2.15}
\end{aligned}$$

The beauty of the result in Equation 4.2.15 is that, while we believe that there exists some g , *we don't need it to approximate* $\mathbb{E}(\theta \mid y)$. With f -modeling, we do not concern ourselves with an unknown prior $g(\theta)$ (Efron, 2017). Instead, we use the given data and our estimated $f_G(y)$ to arrive at an empirical estimate.

Good and Toulmin estimator

Good and Toulmin sought to answer Corbett's question by applying an f -modeling approach to (4.2.3). First, they executed a Taylor expansion of $(1 - e^{-\theta t})$:

$$\begin{aligned}
&= S \int_0^\infty e^{-\theta} \left(\theta t - \frac{(\theta t)^2}{2!} + \frac{(\theta t)^3}{3!} - \dots \right) g(\theta) \, d\theta \\
&= St \int_0^\infty e^{-\theta} \theta g(\theta) \, d\theta - St^2 \int_0^\infty e^{-\theta} \frac{\theta^2}{2!} g(\theta) \, d\theta + \dots
\end{aligned}$$

Recognizing that $e_x = \mathbb{E}(y_x) = S \int_0^\infty \frac{e^{-\theta} \theta^x}{x!} g(\theta) \, d\theta$, they wrote $u(t) = e_1 t - e_2 t^2 + e_3 t^3 - \dots$. Since the true e_x is unknown, they used the known y_x values as an estimate:

$$\begin{aligned}
\hat{u}_{\text{GT}}(t) &= y_1 t - y_2 t^2 + y_3 t^3 - \dots \\
&= \sum_i^N (-t)^{i+1} y_i \tag{4.2.16}
\end{aligned}$$

The Good and Toulmin estimator $\hat{u}_{\text{GT}}(t)$ allows us to answer Corbett's question without establishing a prior distribution g . Applying $t = 0.5$ and the known y_x values,

we have

$$\begin{aligned}\hat{u}_{\text{GT}}(0.5) &= 118(0.5) - 74(.5)^2 + 44(.5)^3 - \dots \\ &= 45.2\end{aligned}\tag{4.2.17}$$

Euler's transformation

While Good and Toulmin (1956) showed that $u_{\text{GT}}(t)$ is unbiased for all $t \leq 1$ and thus proves a good estimator for Corbett's question, it breaks down for all $t > 1$ (Efron and Thisted, 1976; Orlitsky et al., 2016). Efron and Thisted (1976) rectified this problem by applying Euler's transformation to the series in (4.2.16), letting $t = \frac{u}{2-u}$:

$$\begin{aligned}\hat{u}_{\text{GT}}(t) &= y_1t - y_2t^2 + y_3t^3 - \dots \\ \left[\frac{1}{2}(1+t)\right] \hat{u}_{\text{GT}}(t) &= \left[\frac{1}{2}(1+t)\right] [y_1t - y_2t^2 + y_3t^3 - \dots] \\ &= \frac{1}{2} [y_1t - y_2t^2 + y_3t^3 - y_4t^4 + \dots + y_1t^2 - y_2t^3 + y_3t^4 - \dots] \\ &= \frac{1}{2}y_1t + \frac{1}{2}t [(y_1 - y_2)t - (y_2 - y_3)t^2 + (y_3 - y_4)t^3 - \dots] \\ \hat{u}_{\text{GT}}(u) &= \frac{1}{2}y_1u + \frac{1}{2}u [r_1(t)]\end{aligned}\tag{4.2.18}$$

Now, we repeat the process on $r_1(t)$:

$$\begin{aligned}\left[\frac{1}{2}(1+t)\right] r_1(t) &= \left[\frac{1}{2}(1+t)\right] [(y_1 - y_2)t - (y_2 - y_3)t^2 + (y_3 - y_4)t^3 - \dots] \\ &= \frac{1}{2} [(y_1 - y_2)t - (y_2 - y_3)t^2 + (y_3 - y_4)t^3 - \dots + (y_1 - y_2)t^2 - (y_2 - y_3)t^3 + \dots] \\ r_1(u) &= \frac{1}{2}(y_1 - y_2)u + \frac{1}{2}u [\{(y_1 - y_2) - (y_2 - y_3)\}t - \{(y_2 - y_3) - (y_3 - y_4)\}t^2] \\ &= \frac{1}{2}(y_1 - y_2)u + \frac{1}{2}u [r_2(t)]\end{aligned}$$

Substituting $r_1(u)$ back into (4.2.18) gives us

$$\begin{aligned}\hat{u}_{GT}(u) &= \frac{1}{2}y_1u + \frac{1}{2}u\left[\frac{1}{2}(y_1 - y_2)u + \frac{1}{2}u[r_2(t)]\right] \\ &= \frac{1}{2}y_1u + \frac{1}{4}(y_1 - y_2)u^2 + \frac{1}{4}u^2[r_2(t)]\end{aligned}$$

If we continue in this process, we will get the relation from Efron and Thisted (1976):

$$\sum_{i=1}^{\infty} (-1)^{i+1} y_i t^i = \sum_{j=1}^{\infty} \mathcal{E}_j u^j, \quad (4.2.19)$$

where

$$\mathcal{E}_j = \sum_{i=1}^j \binom{j-1}{i-1} \frac{(-1)^{i+1}}{2^j} y_i \quad (4.2.20)$$

Through Euler's transformation, we have created a situation where $\hat{u}_{GT}(y) = \hat{u}_{GT}(u)$ as long as both limits exist. Another benefit of Euler's transformation is that we can use partial sums to estimate the summand. As we continually apply the transformation to our series, the 'differencing of differences' pattern will continue and force the remainder portion $r_p(t)$ to become smaller. Since all of our y_x are positive, we will see a much faster convergence of the partial sums of $\hat{u}_{GT}(u)$.

Next, given a truncation point y_0 and $Z \sim \text{Binomial}(y_0, \frac{1}{1+t})$ Efron and Thisted (1976) derived a new estimator $u_{ET}(t)$ by substituting (4.2.20) into (4.2.19):

$$\begin{aligned}
u_{ET} &= \sum_{j=1}^{y_0} u^j \sum_{i=1}^j \binom{j-1}{i-1} \frac{(-1)^{i+1}}{2^j} y_i \\
&= \sum_{j=1}^{y_0} \left(\frac{t}{1+t} \right)^j \sum_{i=1}^j \binom{j-1}{i-1} (-1)^{i+1} y_i \\
&= \left(\frac{t}{1+t} \right)^1 (-1)^2 y_1 + \left(\frac{t}{1+t} \right)^2 ((-1)^2 y_1 + (-1)^3 y_2) \\
&\quad + \left(\frac{t}{1+t} \right)^3 ((-1)^2 y_1 + 2(-1)^3 y_2 + (-1)^4 y_2) \dots \\
&= (-1)^2 y_1 t^1 \left(\frac{1}{1+t} + \frac{t}{(1+t)^2} + \dots + \frac{t^{y_0-1}}{(1+t)^{y_0}} \right) \\
&\quad + (-1)^3 y_2 t^2 \left(\frac{1}{(1+t)^2} + \frac{t}{(1+t)^3} + \dots + \frac{t^{y_0-2}}{(1+t)^{y_0}} \right) + \dots \\
&\quad + (-1)^{y_0+1} y_{y_0} t^{y_0} \left(\frac{1}{(1+t)^{y_0}} \right) \\
&= (-1)^2 y_1 t^1 \mathcal{P}(Z \geq 1) + (-1)^3 y_2 t^2 \mathcal{P}(Z \geq 2) + \dots + (-1)^{y_0+1} y_{y_0} t^{y_0} \mathcal{P}(Z \geq y_0) \\
&= \sum_{i=1}^{y_0} y_i h_i, \tag{4.2.21}
\end{aligned}$$

where

$$h_i = -(-t)^i \mathcal{P} \left(\text{Bin} \left(y_0, \frac{1}{1+t} \right) \geq i \right). \tag{4.2.22}$$

for $i = 1, \dots, y_0$ and zero otherwise. In addition to truncating the number of terms, u_{ET} incorporates a binomial smoothing parameter that dampens the potential rapid growth of $(-t)^i$ as t gets large (Efron and Thisted, 1976; Orlitsky et al., 2016). While Efron and Thisted (1976) demonstrated the effectiveness of such an approach when applied to the Shakespeare’s vocabulary, it did not provide clear guidelines on selecting the optimal amount of y values. The authors showed that $y_0 = 9$ is an appropriate truncation point for the Shakespeare data using two approaches. First, they calculated the first 20 values of (4.2.20) and observed that they are positive for $y = 1, \dots, 9$ and negative for the remaining values. Furthermore, all of the negative

values where within a single standard deviation of zero thus do not have a significant impact on the estimate. The second approach compared the bias and variance of estimates generated with $y_0 = 9$ and $y_0 = 19$, shown in Table 4.3. While the bias for the estimate at $y_0 = 19$ had a smaller range, the variance of the estimate was extremely large.

Table 4.3: Comparison of the bias and variance between truncation points $y_0 = 9$ and $y_0 = 19$. For the Bias range, y_+ is the sum of all the y values up to the truncation point y_0 .

	$y_0 = 9$	$y_0 = 19$
\hat{u}_{ET}	45188	53867
Bias range	$(-4.24y_+, 0.31y_+)$	$(-1.64y_+, 0.15y_+)$
Standard Deviation	3994	702566

While this analysis supports their selection of y_0 , Efron and Thisted (1976) admitted that their approach was neither ideal nor provided a closed-form solution for determining the truncation point. Orlitsky et al. (2016) addressed this concern and derived a closed-form solution for determining the appropriate and most efficient number of y values to use in missing species estimation. In addition, they also improved on the estimator in Efron and Thisted (1976) and Thisted and Efron (1987). The authors showed that adjusting the probability parameter in the binomial smoothing of (4.2.22) from $p = \frac{1}{1+t}$ to $p = \frac{2}{2+t}$ improves the speed of the estimate's convergence. In addition, Orlitsky et al. (2016) also showed they can achieve this convergence by using fewer y terms.

Extension of DS priors to discrete DS distributions

We have previously described some of the primary approaches to f -modeling for the missing species problem. In all of these approaches, they choose to bypass the prior distribution g and the challenges it brings by nature of working in the θ domain. While our application of the $\text{DS}(G, m)$ prior to g -modeling showed some promise, it

included the challenge of selecting an appropriate g . This challenge becomes more evident, and complicated, when dealing with heavily skewed data present in missing species estimation. Now, the critical question we ask is can we utilize the theory of the DS prior to derive an f -modeling approach? Furthermore, can we build upon the results of Efron and Thisted (1976) and Orlitsky et al. (2016) that allows us to gain optimal results using both truncation and binomial smoothing?

If we choose to take an f -modeling approach, though, we would need to model in the x domain not the θ domain. We will use similar logic as in Chapter 3, only this instance extends the application to discrete probability distributions.

Given $p(x)$, the true mass function based on the given frequencies y_x where $x \in \mathcal{X}$, the set of integer values that governs the number of times a species is observed. Let $g(x)$ be the estimated mass function. We can find an estimated *DS Distribution* $\hat{p}(x)$:

$$\begin{aligned}\hat{p} &= g(x) \times \frac{p}{\bar{p}} \\ &= g(x) \hat{d}[G(x)]\end{aligned}$$

As with the DS prior, the DS distribution consists of a parametric $g(x)$ that represents a probability mass function and the nonparametric $\hat{d}[G(x)]$, the uncertainty function. The critical parameters of our are the LP-Fourier coefficients. As detailed in Chapter 3, instead of finding an expected proxy, we can directly estimate the LP means using their empirical counterpart $\widehat{\text{LP}}[j; G, \Pi] = k^{-1} \sum_{i=1}^k T_j(\theta_i; G)$.

Analysis of Butterfly Data

We first applied the DS Distribution to the Butterfly data. In this case, we chose to conduct the analysis using three different starting distributions for g : the uniform, a zipfian approximation (zipf), and multinomial distribution. We deliberately chose the uniform distribution to test the updated algorithm; we know that the true p was

Table 4.4: Comparison of results using the DS-distribution algorithm with binomial smoothing against other f and g modeling techniques.

	Butterfly $t = 0.50$	Shakespeare $t = 1.00$	‘Shall I Die?’ $t = 0.000485$
DS Distribution with zip	53.80	11394.65	6.97
DS Distribution with Multinomial	48.02	11520.10	6.99
Nonparametric Deconvolution	49.02	11530.43	6.99
Fishers Parametric	46.95	11489.99	6.97
Good-Toulmin	45.17	11486.00	6.97
Efron-Thisted	44.97	11441.29	6.97
Smoothed Good-Toulmin	44.80	11450.60	6.97

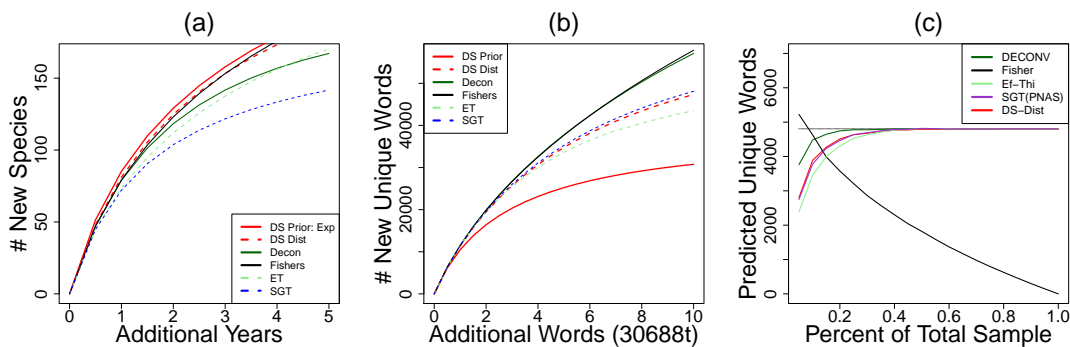


Figure 4.8: Prediction plots that compare results of DS Prior (g -modeling) with DS Distribution (f -modeling) for (a) butterfly and (b) Shakespeare’s total lexicon. Panel (c) demonstrates the effectiveness of DS Distribution in using various percentages of sample size to predict the total words in Shakespeare’s Hamlet.

not evenly distributed among the mass points; if the algorithm works then we would expect some form of correction. After applying the updated algorithm, we found the zipf and multinomial distribution did not require correction. The uniform, though, did require correction and has the following LP representation:

$$\begin{aligned}
 d[G(y)] = & 1 - 0.853T_1(y, G) + 0.601T_2(y, G) - 0.458T_3(y, G) + \\
 & 0.274T_4(y, G) - 0.243T_5(y, G) + 0.123T_6(y, G) \quad (4.2.23)
 \end{aligned}$$

For the uniform distribution, we observed that the corrections generated by the algorithm created a g that was very similar to both the zipf and multinomial. We

used the zipf and multinomial distributions to predict the number of new species at $t = 0.5$, shown in the first column of Table 4.4. In all cases, we see that we are well within the standard error associated with the Good and Toulmin’s f -modeling estimate: 45.2 ± 4.4 (Good and Toulmin, 1956; Efron and Hastie, 2016; Efron, 2017).

Analysis of Shakespeare’s Vocabulary

We now apply the DS distribution to another common application of the missing species, the prediction of an individual’s knowledge of language. In particular, we want to apply our method to the Shakespeare lexicon and compare our results with Efron and Thisted (1976). Our goal is to predict the number of words Shakespeare knew based on the number of words he actually used. His known works contain 884,647 total words with 31,534 unique word types. Table 4.5 shows the total number of word types y used x times.

Table 4.5: The frequencies for the unique word types that Shakespeare used in his works, y_x is the total number of word types y used x times (Efron and Thisted, 1976). Not shown are 846 word types that appear more than 100 times.

x	1	2	3	4	5	6	7	8	9	10
0+	14376	4343	2292	1463	1043	837	638	519	430	364
10+	305	259	242	223	187	181	179	130	127	128
20+	104	105	99	112	93	74	83	76	72	63
30+	73	47	56	59	53	45	34	49	45	52
40+	49	41	30	35	37	21	41	30	28	19
50+	25	19	28	27	31	19	19	22	23	14
60+	30	19	21	18	15	10	15	14	11	16
70+	13	12	10	16	18	11	8	15	12	7
80+	13	12	11	8	10	11	7	12	9	8
90+	4	7	6	7	10	10	15	7	7	5

For the zipf distribution, $d[G(y)]$ had only 4 significant LP coefficients:

$$d[G(y)] = 1 - 0.747T_1(y, G) + 0.257T_2(y, G) - 0.084T_3(y, G) + 0.025T_4(y, G) \quad (4.2.24)$$

The multinomial g , on the other hand, resulted in no significant LP coefficients. We used both models to predict the total number of unique words Shakespeare knew and presented the results in the second column of Table 4.4. The DS Distribution results are comparable with other techniques; from Efron and Hastie (2016) we know that Robbins’ formula gives us 11430 ± 178 for the expected number of distinct new words if we discovered a new work just as large as the sample of his old work (i.e. 884,647 words, $t = 1$). The standard error for the expected number of words $\hat{E}(t)$ is approximated by

$$\widehat{\text{sd}}(t) = \left(\sum_{i=1}^{100} y_x t^{2x} \right)^{\frac{1}{2}} \quad (4.2.25)$$

Both DS estimates fall within a single standard error, indicating that our approximation provides a reasonable and accurate alternative to other methods. We can further validate our results through the poem “Shall I die?” Discovered in 1985, some experts attribute the poem of 429 words to Shakespeare (Efron, 2016). We determine t by dividing the words in the poem by the total canon and determine the number of expected new words; the results are shown in the third column of Table 4.4. Again, our predictions are in line with both current and past methods.

Analysis of Hamlet Data

In this simulation, we will use varied sample sizes from Shakespeare’s *Hamlet* to assess the DS Distribution’s ability to predict a vocabulary size based on a partial text (Orlitsky et al., 2016). The play itself contains $N = 31,999$ total words, where 4,804 of those words are unique. First, the simulation will randomly sample n of the N words without replacement, where $n = pN$ and p is some percentage between 0.05 and 1. For each n , we will use that sample to predict the number of words that are not in the sample, \hat{n} . For each p , we will repeat the simulation for one hundred iterations and average the results. The goal is to have $n + \hat{n} \approx N$. We will predict using Efron’s

nonparametric Bayesian deconvolution (DECONV, a g -modeling technique), Fisher’s parametric approach (4.2.6), Efron-Thisted’s estimator in (4.2.21) and (4.2.22), and the smoothed Good-Toulmin estimator presented in Orlitsky et al. (2016). Essentially, this estimator is the Efron-Thisted estimator except $p = \frac{2}{2+t}$ in (4.2.22).

Panel (c) of Figure 4.8 shows the results of the simulation. First, we look at the two examples of g -modeling in our analysis: DECONV and Fisher’s parametric approach. It is clear that the nonparametric deconvolution result (green) performs best; the truly nonparametric nature can use small sample sizes to appropriately model the data and produce accurate predictions. On the contrary, Fisher’s parametric g -modeling (black) struggles. As the sample size increases, we see a decrease in the actual predicted words. The next three come from f -modeling: Efron-Thisted (light blue), Smoothed Good-Toulmin (purple), and the DS Distribution (red). All three of these models demonstrate the same behavior: the predictive power increases as the sample size increases. The plot shows the DS distribution slightly outperforms the Smoothed Good-Toulmin, but not at a level we would deem significant.

The simulation demonstrates that the purely nonparametric approach greatly outperforms the rest. We acknowledge that three of the f -modeling approaches require only the first few y_i to play a key role in the algorithm due to the thresholding of x_0 . The Bayesian deconvolution method, though, incorporates all of the data which can contribute to the increased predictive power. So why don’t we abandon our f -modeling approaches in favor of the g ? To answer the question, we remind ourselves that the Bayesian deconvolution approach is g -modeling and requires the approximation of a prior distribution before predicting. As shown in our g -modeling discussion, this approximation is not always a simple and straightforward task.

The f -modeling approach bypasses the prior distribution and provides a functional form to predict based only on the given data and a selected smoothing criteria. More importantly, though, we have reinforced the extension of the $DS(G, m)$ construct

from a prior distribution used in hierarchical modeling to one that can model *any* distribution. With this extension, we can now generate appropriate and accurate distributional models for any data. Consequently, these improved models will help improve analysis and inferences in support of decision making.

4.2.3 Key Observations

This section demonstrated the flexibility of gEB by its application to the missing species problem with both g and f modeling. In both cases, the results from the gEB were either comparable or improvements to popular and current algorithms. During g -modeling, we saw the complexities of selecting an appropriate prior distribution. These complexities may prove too daunting for a fledgling data scientist, thus the simple beauty of gEB's use of any prior distribution can assist in simplifying the situation. During the f -modeling, we established a framework by which we can extend the DS prior concept to discrete, unconditional distributions. The missing species problem reinforces both the flexibility and reliability of generalized empirical Bayes.

4.3 Summary

The purpose of this chapter was to not only illustrate the versatility of generalized empirical Bayes, but demonstrate its effectiveness at addressing a diverse selection of problems. First, we took an example that represents a significantly complex structure for modern learning algorithms then systematically demonstrated how the $DS(G, m)$ prior succeeds in capturing the true nature of the data. Next, we showed the flexibility of the structure of the $DS(G, m)$ prior itself and how we can leverage its uncertainty function to find both Bayesian and frequentist solutions to a historical yet relevant problem.

What do both of these applications have in common? While one could argue that a prior distribution g appears in both scenarios, our frequentist approach to the missing species problem is a strong counter example to that suggestion. Before gEB and the $DS(G, m)$ model, these two applications represented distinct problems with solutions tailored to their specific data structure and desired results. This chapter has shown how the $DS(G, m)$ model is more than just a new class of prior distributions in a hierarchical model. It represents a powerful tool for analysts and decision makers to improve insight and inferences from a variety of problem types.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This dissertation provides extensive details into the theory and applications of a radical new data modeling paradigm that unifies Bayesian and frequentist modeling philosophies. We have progressively developed the concepts and principles through a range of examples, spanning application areas such as clinical trials, metrology, insurance, medicine, and ecology, that highlight the core of our approach: an elegant combination of Bayesian thinking (parameter probability where prior knowledge can be encoded) with frequentist goodness-of-fit (evaluation and synthesis of the prior distribution). The foundation of the approach is the $DS(G, m)$ prior distribution which couples a known parametric prior with a nonparametric correction factor. In addition, we showed how to adapt generalized empirical Bayes to a robust five-tiered process to analyze uncertain data that tackles modern concerns of uncertainty and scientific discovery. We provided a new perspective on the science behind combining estimates from distributed data, one that learns a more insightful answer without

gaining access to all available data. Finally, we used the missing species problem to show how the expansion of the uncertainty function d is equally powerful in both Bayesian and frequentist analysis.

While this dissertation introduces several unique and exciting applications of generalized empirical Bayes, it also serves as the inspiration to explore how gEB supports other areas. The next two sections highlight areas both significant and groundbreaking. First, we seek to expand the utility of the $DS(G, m)$ prior with a more traditional Bayesian approach in determining its parameters. Second, there are significant applications for gEB in the field of cybersecurity modeling.

5.2 Moving toward Full Bayesian Analysis

In Chapter 1, we introduced generalized empirical Bayes as a third culture of Empirical Bayes, a philosophy that estimates the unknown prior distribution $\pi(\theta)$ from the observed data (Morris, 1983; Efron, 2017). This philosophy is counter to the more traditional Bayesian approach, Full Bayesian Analysis, which avoids the estimation of $\pi(\theta)$ from the data. Instead, Full Bayesian Analysis places a probability model on the complete set of parameters and analyzes their joint distribution (Gelman et al., 2013). For example, when we consider the rat example introduced in Chapter 2, we have the following model for Full Bayesian Analysis:

$$y_i | \theta_i \sim \text{Binomial}(n_i, \theta_i), \quad i = 1, \dots, k$$

$$\theta_i \sim \pi(\theta) = \text{Beta}(\alpha, \beta)$$

$$\alpha \sim \mathcal{N}(2.4, 5)$$

$$\beta \sim \mathcal{N}(14, 5)$$

In this model, we do not have any fixed parameters. Instead, all parameters in the model are governed by either a prior distribution $\pi(\theta)$ or hyperprior distributions. The hyperprior is a prior distribution placed on the parameters of the prior; in this case, we allow α and β to be governed by a normal distribution centered at their respective MLE estimates with an inflated variance. To estimate the parameters, the next step in Full Bayesian analysis is to run the model through Markov Chain Monte Carlo (MCMC) simulation.

The model parameters are critical to effective and efficient MCMC, which is the critical tool in Full Bayes analysis. While the $DS(G, m)$ behaves like a nonparametric, it is in fact parametric with both the parameters for the selected prior g and the LP means $LP[j; G, \Pi], j = 1 \dots m$. In terms of the previous model, we can use the hyperpriors for α and β but need to identify an appropriate hyperprior for the LP-Fourier coefficients. From Mukhopadhyay (2017), we know that the limiting distribution sample LP means as $k \rightarrow \infty$ is $\mathcal{N}(0, k^{-1})$. Therefore, we can implement the following model for $DS(G, m)$, where $G = \text{Beta}(\alpha, \beta)$:

$$\begin{aligned}
 y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i), \quad i = 1, \dots, k \\
 \theta_i &\sim \pi(\theta) \equiv DS(G, m) = g(\theta) \left[1 + \sum_{j=1}^m LP[j; G, \Pi] \text{Leg}_j(G(\theta)) \right] \\
 \alpha &\sim \mathcal{N}(2.4, 5) \\
 \beta &\sim \mathcal{N}(14, 5) \\
 LP[j; G, \Pi] &\sim \mathcal{N}(0, 5), \quad j = 1, \dots, m
 \end{aligned}$$

where $\text{Leg}_j(u)$ is the j th orthonormalized shifted Legendre polynomial:

$$\begin{aligned}\text{Leg}_1(u) &= \sqrt{12}\left(u - \frac{1}{2}\right) \\ \text{Leg}_2(u) &= \sqrt{5}(6u^2 - 6u + 1) \\ &\vdots\end{aligned}$$

In this model, we follow common Bayesian practice to inflate the variance of the hyperprior distributions to allow the data maximum flexibility within the simulation. When we run MCMC using this model, the model specification will allow the simulation to discover the appropriate parameters and enable follow-on inference using the posterior. The MCMC process generates a standard error for each estimate, which in the case of the LP-Fourier coefficients allows the identification of significant terms. We used the rat data to motivate the discussion about Full Bayesian analysis for $\text{DS}(G, m)$ priors, but the beauty of implementing MCMC for the $\text{DS}(G, m)$ prior lies in the capability to ‘break free’ of the conjugate requirement. Successful implementation of the $\text{DS}(G, m)$ into MCMC permits the use of any prior distribution and increases the utility of gEB.

Critical to this implementation are the orthonormalized shifted Legendre polynomials. Many MCMC simulation software packages include a variety of probability distributions, but none include the $\text{DS}(G, m)$ prior nor the appropriate Legendre polynomials. In order to successfully use a $\text{DS}(G, m)$ prior in MCMC, we require a software package that allows custom distributions and the capability to adapt and tune the MCMC algorithm to properly navigate the $\text{DS}(G, m)$ prior. While there are some applicable candidates for and existing packages, we are also pursuing an option where we code our own algorithm.

5.3 Modeling Attack Types for Improved Cyber-security

One of the most prevalent types of cyber attacks are those that result in loss of personal or proprietary data. Known as data breaches, these seemingly non-destructive attacks withdraw valuable data for profit (Maillart and Sornette, 2010; Edwards, 2016). These breaches typically target large organizations or companies and can have traumatic effects on these organizations and consumers (Wheatley et al., 2016; Xu et al., 2018). To reduce or eliminate the impact of these events, researchers want accurate predictive models, particularly as it pertains to the size of the next breach and when it will occur. Current work aims to improve predictions by developing insight into the size, when they occur, the inter-arrival time of breaches, and the length of time to investigate and remediate breaches (Maillart and Sornette, 2010; Edwards et al., 2016; Wheatley et al., 2016; Xu et al., 2018). While all of this work will help improve our understanding of data breaches, the majority is based on using data from *known* breaches (Edwards et al., 2016; Xu et al., 2018). The challenge, though, is that not all data breaches are publically disclosed.

Our framework can lend assistance in improving the models and understanding of data breaches, but has potential to provide breakthroughs in understanding how attackers *are successful* in creating these breaches. MITRE (2018) outlines different attack frameworks that adversarial hackers use to gain access to an enterprise network. While it gives a generalization of the frameworks, the reality is hackers are individuals that approach each attack with unique and varied experiences. The different forms of attack and hacker individualism introduces heterogeneity within each breach attempt. Furthermore, more sophisticated hackers will attempt to disguise their actions as those deemed permissible by the network’s current security systems. Not only do we have heterogeneity between different skill levels of hackers, but we also have heterogeneity

between hackers and those individuals with access to the system. The multiple cases of heterogeneity provides an opportunity for the $DS(G, m)$ model.

For a specific attack technique A_j let $j = 1$ is the designation for the first attack type, $j = 2$ is the second, and so on. For A_1 , we would collect both host and network based data on individual attackers y_i , where $i = 1, \dots, k$. Using gEB, we would generate the DS prior distribution for A_1 ; this distribution will smartly combine the information from various hackers into one more representative summary of the attack. We continue this process for all particular attack types until we have models for A_j , with $j \geq 1$. At this point, we have a collection of local, ‘partition-level’ estimates for different attack types...but how can these individualized models help identify a breach? We need to apply another iteration of combining with gEB at the enterprise level.

Due to the variety of attacks available, trying to determine if the available data matches one of the many local models is time consuming and inefficient. We need to incorporate these local models into a *global model* that quickly identifies an attack. More importantly, because the global model is built using multiple local models, this form of anomaly detection can also classify the type of attack. While identification is key, classification of the type of attack can help organizations and IT professionals quickly address any issues caused by the attack (Kuypers et al., 2016).

Implementation of this form of anomaly detection and classification requires extensive data collection and efficient modeling. Currently, the absence of reliable data is a significant challenge to statistical efforts in cybersecurity (Edwards et al., 2016; Xu and Hua, 2017; Xu et al., 2018). There are opportunities to leverage activities hosted by the Army Cyber Institute to begin generation of the required data. In terms of efficient modeling, the global model should be able to run as close to real-time as possible for any size of organization. We should be able to scale this model so that it can work efficiently and quickly for any environment.

BIBLIOGRAPHY

- Bailey, D. (2018), “Why OUTLIERS are good for science,” *Significance*, 15, 14–19.
- Bailey, D. C. (2017), “Not Normal: the uncertainties of scientific measurements,” *Royal Society open science*, 4, 160600.
- Beckett, L. and Diaconis, P. (1994), “Spectral analysis for discrete longitudinal data,” *Advances in Mathematics*, 103, 107–128.
- Berger, J. and Berliner, L. M. (1986), “Robust Bayes and empirical Bayes analysis with ε -contaminated priors,” *The Annals of Statistics*, 461–486.
- Berger, J. O. (1994), “An overview of robust Bayesian analysis (with discussion),” *Test*, 3, 5–124.
- (2000), “Bayesian analysis: A look at today and thoughts of tomorrow,” *Journal of the American Statistical Association*, 95, 1269–1276.
- Brillinger, D. R. (2002), “John W. Tukey: his life and professional contributions,” *Annals of Statistics*, 1535–1575.
- Brimacombe, M. (2014), “Genomic aggregation effects and Simpson’s paradox,” *Open Access Medical Statistics*.
- Bunge, J. and Fitzpatrick, M. (1993), “Estimating the number of species: a review,” *Journal of the American Statistical Association*, 88, 364–373.

- Casella, G. and Berger, R. L. (2002), *Statistical inference*, vol. 2, Duxbury Pacific Grove, CA.
- Chao, A. (2005), “Species estimation and applications,” *Encyclopedia of statistical sciences*.
- Chen, G., Gott III, J. R., and Ratra, B. (2003), “Non-Gaussian Error Distribution of Hubble Constant Measurements,” *Publications of the Astronomical Society of the Pacific*, 115, 1269.
- Cox, D. and Efron, B. (2017), “Statistical thinking for 21st century scientists,” *Science Advances*, 3, e1700768.
- Cox, D. R. (1990), “Comment: The 1988 Wald Memorial Lectures: The Present Position in Bayesian Statistics,” *Statistical Science*, 5, 76–78.
- Crandall, S., Houston, S., and Ratra, B. (2015), “Non-Gaussian error distribution of 7 Li abundance measurements,” *Modern Physics Letters A*, 30, 1550123.
- De Blok, W., McGaugh, S. S., and Rubin, V. C. (2001), “High-resolution rotation curves of low surface brightness galaxies. II. Mass models,” *The Astronomical Journal*, 122, 2396.
- Dempster, A. P. (1975), “A subjectivist look at robustness,” *Bulletin of the International Statistical Institute*, 46, 349–374.
- Dempster, A. P., Selwyn, M. R., and Weeks, B. J. (1983), “Combining historical and randomized controls for assessing trends in proportions,” *Journal of the American Statistical Association*, 78, 221–227.
- Edwards, B. (2016), “Evidence-based Cybersecurity: Data-driven and Abstract Models,” *Dissertation*.

- Edwards, B., Hofmeyr, S., and Forrest, S. (2016), “Hype and heavy tails: A closer look at data breaches,” *Journal of Cybersecurity*, 2, 3–14.
- Efron, B. (1975), “Defining the curvature of a statistical problem (with applications to second order efficiency),” *The Annals of Statistics*, 1189–1242.
- (1986), “Why isn’t everyone a Bayesian?” *The American Statistician*, 40, 1–5.
- (1996), “Empirical Bayes methods for combining likelihoods,” *Journal of the American Statistical Association*, 91, 538–550.
- (2003), “Robbins, empirical Bayes and microarrays,” *The Annals of Statistics*, 31, 366–378.
- (2012), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press.
- (2013), *Empirical Bayes modeling, computation and accuracy*, Division of Biostatistics, STANFORD University.
- (2016), “Empirical Bayes deconvolution estimates,” *Biometrika*, 103, 1–20.
- (2017), “Bayes, Oracle Bayes, and Empirical Bayes,” *Technical Report No.2017-12*.
- Efron, B. and Hastie, T. (2016), *Computer Age Statistical Inference*, vol. 5, Cambridge University Press.
- Efron, B. and Morris, C. (1973), “Stein’s estimation rule and its competitorsan empirical Bayes approach,” *Journal of the American Statistical Association*, 68, 117–130.
- (1975), “Data analysis using Stein’s estimator and its generalizations,” *Journal of the American Statistical Association*, 70, 311–319.
- Efron, B. and Morris, C. N. (1977), *Stein’s paradox in statistics*, WH Freeman.

- Efron, B. and Thisted, R. (1976), “Estimating the number of unseen species: How many words did Shakespeare know?” *Biometrika*, 63, 435–447.
- Ferguson, T. S. (1973a), “A Bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, 209–230.
- (1973b), “A Bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, 209–230.
- Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943), “The relation between the number of species and the number of individuals in a random sample of an animal population,” *The Journal of Animal Ecology*, 42–58.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis, Third Edition*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.
- Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008), “A weakly informative default prior distribution for logistic and other regression models,” *The Annals of Applied Statistics*, 1360–1383.
- Gelman, A., Meng, X.-L., and Stern, H. (1996), “Posterior predictive assessment of model fitness via realized discrepancies,” *Statistica Sinica*, 733–760.
- Gelman, A., Simpson, D., and Betancourt, M. (2017), “The prior can often only be understood in the context of the likelihood,” *Entropy*, 19, 555.
- Good, I. (1992), “The Bayes/non-Bayes compromise: A brief review,” *Journal of the American Statistical Association*, 87, 597–606.
- Good, I. and Toulmin, G. (1956), “The number of new species, and the increase in population coverage, when a sample is increased,” *Biometrika*, 43, 45–63.

- Good, I. J. (1983), “The philosophy of exploratory data analysis,” *Philosophy of science*, 50, 283–295.
- Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016), “What does research reproducibility mean?” *Science translational medicine*, 8, 341ps12–341ps12.
- Gu, J. and Koenker, R. (2016), “On a problem of Robbins,” *International Statistical Review*, 84, 224–244.
- Hedges, L. V. (1983), “A random effects model for effect sizes.” *Psychological Bulletin*, 93, 388.
- Hedges, L. V. and Olkin, I. (1985), “Statistical methods for meta-analysis,” .
- Higgins, J. P. and Thompson, S. G. (2002), “Quantifying heterogeneity in a meta-analysis,” *Statistics in medicine*, 21, 1539–1558.
- Hinshaw, G., Spergel, D., Verde, L., Hill, R., Meyer, S., Barnes, C., Bennett, C., Halpern, M., Jarosik, N., Kogut, A., et al. (2003), “First-Year Wilkinson Microwave Anisotropy Probe (WMAP)* Observations: The Angular Power Spectrum,” *The Astrophysical Journal Supplement Series*, 148, 135.
- Huchra, J. (2008), “The Hubble Constant,” [Online; accessed 31-October-2018].
- Ioannidis, J. P. (2005), “Why most published research findings are false,” *PLoS medicine*, 2, e124.
- James, W. and Stein, C. (1961), “Estimation with quadratic loss,” in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379.
- Jeffreys, H. (1938), “The law of error and the combination of observations,” *Phil. Trans. R. Soc. Lond. A*, 237, 231–271.

- Kafadar, K. et al. (2003), “John Tukey and robustness,” *Statistical Science*, 18, 319–331.
- Kiefer, J. and Wolfowitz, J. (1956a), “Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters,” *The Annals of Mathematical Statistics*, 887–906.
- (1956b), “Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters,” *The Annals of Mathematical Statistics*, 887–906.
- Kievit, R., Frankenhuis, W. E., Waldorp, L., and Borsboom, D. (2013), “Simpson’s paradox in psychological science: a practical guide,” *Frontiers in psychology*, 4, 513.
- Koenker, R. and Gu, J. (2017), “REBayes: Empirical Bayes Mixture Methods in R,” *Journal of Statistical Software*, 82.
- Koenker, R. and Mizera, I. (2014), “Convex optimization, shape constraints, compound decisions, and empirical Bayes rules,” *Journal of the American Statistical Association*, 109, 674–685.
- Kuypers, M. A., Maillart, T., and Pate-Cornell, E. (2016), “An empirical analysis of cyber security incidents at a large organization,” *Department of Management Science and Engineering, Stanford University, School of Information, UC Berkeley*, http://fsi.stanford.edu/sites/default/files/kuypersweis_v7.pdf, accessed July, 30.
- Lehmann, E. L. and Casella, G. (2006), *Theory of point estimation*, Springer Science & Business Media.
- Liu, J. S. (1996), “Nonparametric hierarchical Bayes via sequential imputations,” *The Annals of Statistics*, 911–930.

- Liu, Q. and Ihler, A. T. (2014), “Distributed estimation, information loss and exponential families,” in *Advances in Neural Information Processing Systems*, pp. 1098–1106.
- Lukoianova, T. and Rubin, V. L. (2014), “Veracity roadmap: Is big data objective, truthful and credible?” *Advances In Classification Research Online*, 24.
- Maillart, T. and Sornette, D. (2010), “Heavy-tailed distribution of cyber-risks,” *The European Physical Journal B*, 75, 357–364.
- Marin-Martinez, F. and Sanchez-Meca, J. (2010), “Weighting by inverse variance or by sample size in random-effects meta-analysis,” *Educational and Psychological Measurement*, 70, 56–73.
- Maritz, J. (1969), “Empirical Bayes estimation for the Poisson distribution,” *Biometrika*, 56, 349–359.
- Martz, H. and Lian, M. (1974), “Empirical Bayes estimation of the binomial parameter,” *Biometrika*, 61, 517–523.
- McDonald, R., Mohri, M., Silberman, N., Walker, D., and Mann, G. S. (2009), “Efficient large-scale distributed training of conditional maximum entropy models,” in *Advances in Neural Information Processing Systems*, pp. 1231–1239.
- Minsker, S. and Strawn, N. (2017), “Distributed Statistical Estimation and Rates of Convergence in Normal Approximation,” *arXiv preprint arXiv:1704.02658*.
- MITRE (2018), “MITRE Adversarial Tactics, Techniques and Common Knowledge (ATT & CK),” [Online; accessed 31-August-2018].
- Morris, C. N. (1983), “Parametric empirical Bayes inference: theory and applications,” *Journal of the American Statistical Association*, 78, 47–55.

- Mukhopadhyay, S. (2017), “Large-scale mode identification and data-driven sciences,” *Electronic Journal of Statistics*, 11, 215–240.
- Mukhopadhyay, S. and Fletcher, D. (2018a), *BayesGOF: Bayesian Modeling via Frequentist Goodness-of-Fit*, R package version 5.2.
- (2018b), “Generalized Empirical Bayes Modeling via Frequentist Goodness of Fit,” *Nature’s Scientific Reports*, 8, 9983.
- Mukhopadhyay, S. and Parzen, E. (2014), “LP approach to statistical modeling,” *arXiv preprint arXiv:1405.2601*.
- Norberg, R. (1989), “Experience rating in group life insurance,” *Scandinavian Actuarial Journal*, 1989, 194–224.
- O’Hagan, A. et al. (2006), “Science, subjectivity and software (comment on articles by Berger and by Goldstein),” *Bayesian Analysis*, 1, 445–450.
- Orlitsky, A., Suresh, A. T., and Wu, Y. (2016), “Optimal prediction of the number of unseen species,” *Proceedings of the National Academy of Sciences*, 113, 13283–13288.
- Possolo, A. (2013), “Five examples of assessment and expression of measurement uncertainty,” *Applied Stochastic Models in Business and Industry*, 29, 1–18.
- Reimer, A. P. and Madigan, E. A. (2018), “Veracity in big data: How good is good enough,” *Health informatics journal*, 1460458217744369.
- Robbins, H. (1951), “Asymptotically subminimax solutions of compound statistical decision problems,” in *Proceedings of the Second Berkley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, vol. I, pp. 131–149.

- (1956), “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*, pp. 157–164.
- (1980), “An empirical Bayes estimation problem,” *Proceedings of the National Academy of Sciences*, 77, 6988–6989.
- Rosenblatt, J. D. and Nadler, B. (2016), “On the optimality of averaging in distributed statistical learning,” *Information and Inference: A Journal of the IMA*, 5, 379–404.
- Rubin, D. B. (1984), “Bayesianly justifiable and relevant frequency calculations for the applied statistician,” *The Annals of Statistics*, 12, 1151–1172.
- Rukhin, A. L. and Vangel, M. G. (1998), “Estimation of a common mean and weighted means statistics,” *Journal of the American Statistical Association*, 93, 303–308.
- Sacks, H. S., Chalmers, T. C., Blum, A. L., Berrier, J., and Pagano, D. (1990), “Endoscopic hemostasis: an effective therapy for bleeding peptic ulcers,” *Journal of the American Medical Association*, 264, 494–499.
- Sims, C. (2010), “Understanding non-Bayesians,” *Unpublished chapter, Department of Economics, Princeton University*.
- Sivaganesan, S. and Berger, J. (1993), “Robust Bayesian analysis of the binomial empirical Bayes problem,” *Canadian Journal of Statistics*, 21, 107–119.
- Sobers, R. (2012), “5 Things You Should Know About Big Data,” [Online; accessed 06-November-2018].
- Stein, C. (1955), “Inadmissibility of the usual estimator for the mean of a multivariate normal distribution,” *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 197–206.

- Stigler, S. M. (1982), “Thomas Bayes’s Bayesian inference,” *Journal of the Royal Statistical Society. Series A (General)*, 125, 250–258.
- Tarone, R. E. (1982), “The use of historical control information in testing for a trend in proportions,” *Biometrics*, 38, 215–220.
- Thisted, R. and Efron, B. (1987), “Did Shakespeare write a newly-discovered poem?” *Biometrika*, 74, 445–455.
- Toman, B. and Possolo, A. (2009), “Laboratory effects models for interlaboratory comparisons,” *Accreditation and Quality Assurance*, 14, 553–563.
- Tukey, J. W. (1948), “Approximate Weights,” *Ann. Math. Statist.*, 19, 91–92.
- (1972), “Exploratory data analysis: as part of a larger whole,” in *Proceedings of the 18th Conference on Design of Experiments in Army Research and Development I. Washington, DC*, vol. 1010.
- Valiant, P. and Valiant, G. (2013), “Estimating the unseen: improved estimators for entropy and other properties,” in *Advances in Neural Information Processing Systems*, pp. 2157–2165.
- Wang, Y. (2007), “On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 185–198.
- Wheatley, S., Maillart, T., and Sornette, D. (2016), “The extreme risk of personal data breaches and the erosion of privacy,” *The European Physical Journal B*, 89, 7.
- Willie, S. and Berman, S. (1995), “Ninth round intercomparison for trace metals in marine sediments and biological tissues,” *NRC/NOAA*.

- Wu, Y. and Yang, P. (2015), “Chebyshev polynomials, moment matching, and optimal estimation of the unseen,” *arXiv preprint arXiv:1504.01227*.
- Xi, R., Lin, N., Chen, Y., and Kim, Y. (2012), “Compression and aggregation of Bayesian estimates for data intensive computing,” *Knowledge and information systems*, 33, 191–212.
- Xie, X., Kou, S., and Brown, L. (2016), “Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance,” *Annals of Statistics*, 44, 564.
- Xu, M. and Hua, L. (2017), “Cybersecurity Insurance: Modeling and Pricing,” Tech. rep., Research Report for the Society of Actuaries. <https://www.soa.org/Files/Research/Projects/cybersecurity-insurancereport.pdf>.
- Xu, M., Schweitzer, K. M., Bateman, R. M., and Xu, S. (2018), “Modeling and Predicting Cyber Hacking Breaches,” *IEEE Transactions on Information Forensics and Security*, 13, 2856–2871.
- Young-Xu, Y. and Chan, K. A. (2008), “Pooling overdispersed binomial data to estimate event rate,” *BMC Medical Research Methodology*, 8, 58.
- Zhang, Y., Wainwright, M. J., and Duchi, J. C. (2012), “Communication-efficient algorithms for statistical optimization,” in *Advances in Neural Information Processing Systems*, pp. 1502–1510.

APPENDIX

R CODE FOR CHAPTER 2

This appendix provides the R-code to use the `BayesGOF` package to recreate select results from Chapter 2. The package is freely available for download from the CRAN website (Mukhopadhyay and Fletcher, 2018a) or directly through R.

```
1 library(BayesGOF)
2 ###
3 # Fig 2.2 (PLOT OF LP FUNCTIONS)
4 ###
5 # Determine polynomials
6 theta.grid <- seq(0,1, length.out = 500)
7 basis.fun.1 <- gLP.basis(theta.grid, c(.5,.5), 4, con.prior = "Beta")
8 basis.fun.2 <- gLP.basis(theta.grid, c(1,1), 4, con.prior = "Beta")
9 basis.fun.3 <- gLP.basis(theta.grid, c(3,4), 4, con.prior = "Beta")
10 # Generate plots
11 dev.new(height = 5, width = 12)
12 par(mfrow = c(1,3))
13 plot(theta.grid, basis.fun.1[,1], type = "l", lwd = 2,
14       xlab = expression(theta), ylab = "", main = "(a)")
15 lines(theta.grid, basis.fun.1[,2], lwd = 2, col = "red")
16 lines(theta.grid, basis.fun.1[,3], lwd = 2, col = "blue")
17 lines(theta.grid, basis.fun.1[,4], lwd = 2, col = "green")
```

```

18 plot(theta.grid, basis.fun.2[,1], type = "l", lwd = 2,
19       xlab = expression(theta), ylab = "", main = "(b)")
20 lines(theta.grid, basis.fun.2[,2], lwd = 2, col = "red")
21 lines(theta.grid, basis.fun.2[,3], lwd = 2, col = "blue")
22 lines(theta.grid, basis.fun.2[,4], lwd = 2, col = "green")
23 plot(theta.grid, basis.fun.3[,1], type = "l", lwd = 2,
24       xlab = expression(theta), ylab = "", main = "(c)")
25 lines(theta.grid, basis.fun.3[,2], lwd = 2, col = "red")
26 lines(theta.grid, basis.fun.3[,3], lwd = 2, col = "blue")
27 lines(theta.grid, basis.fun.3[,4], lwd = 2, col = "green")
28 ####
29 # Fig 2.4
30 ####
31 # Load data
32 data(rat); data(terb); data(ulcer); data(tacks)
33 # Find g
34 rat.g <- gMLE.bb(rat$y, rat$n)$estimate
35 terb.g <- gMLE.bb(terb$y, terb$n)$estimate
36 ulc.g <- gMLE.nn(ulcer$y, ulcer$se)$estimate
37 tack.g <- gMLE.bb(tacks$y, tacks$n)$estimate
38 # Find DS prior
39 rat.ds <- DS.prior(rat, max.m = 4, g.par = rat.g,
40                  family = "Binomial")
41 terb.ds <- DS.prior(terb, max.m = 8, g.par = terb.g,
42                  family = "Binomial")
43 ulc.ds <- DS.prior(ulcer, max.m = 8, g.par = ulc.g,
44                  family = "Normal")
45 tack.ds <- DS.prior(tacks, max.m = 6, g.par = tack.g,
46                  family = "Binomial")
47 # Generate plots of U function
48 dev.new(height = 5, width = 15)
49 par(mfrow = c(1,3))
50 plot(rat.ds, plot.type = "Ufunc", main = "(a)")

```

```

51 plot(terb.ds, plot.type = "Ufunc", main = "(b)")
52 plot(tack.ds, plot.type = "Ufunc", main = "(c)")
53 ###
54 # Fig 2.5 (MaxEnt Comparison)
55 ###
56 # Galaxy and Rat DS (MaxEnt)
57 data(galaxy)
58 gal.g <- gMLE.nn(galaxy$y, galaxy$se)$estimate
59 gal.ds <- DS.prior(galaxy, max.m = 6, g.par = gal.g,
60                   family = "Normal")
61 gal.ME <- DS.prior(galaxy, max.m = 6, g.par = gal.g,
62                   family = "Normal", LP.type = "MaxEnt")
63 rat.ME <- DS.prior(rat, max.m = 4, g.par = rat.g,
64                   family = "Binomial", LP.type = "MaxEnt")
65 # Generate plots
66 dev.new(height = 5, width = 10)
67 par(mfrow = c(1,2))
68 plot(rat.ds$prior.fit$theta.vals, rat.ds$prior.fit$ds.prior,
69       col = "red", type = "l", xlim = c(0,.4),
70       main = "(a)")
71 lines(rat.ME$prior.fit$theta.vals, rat.ME$prior.fit$ds.prior,
72       col = "darkgreen")
73 legend("topright", c("DS:L2", "DS:MaxEnt"),
74       col = c("red", "darkgreen"), lwd = 4)
75 plot(gal.ME$prior.fit$theta.vals, gal.ME$prior.fit$ds.prior,
76       col = "darkgreen", type = "l", xlim = c(-100,250),
77       main = "(b)")
78 lines(gal.ds$prior.fit$theta.vals, gal.ds$prior.fit$ds.prior,
79       col = "red")
80 legend("topright", c("DS:L2", "DS:MaxEnt"),
81       col = c("red", "darkgreen"), lwd = 4)
82 ###
83 # Fig 2.6 (Deviance Plots)

```

```

84 ####
85 # Ship and surgical node DS
86 data(ship); data(surg)
87 g.surg <- gMLE.bb(surg$y, surg$n)$estimate
88 ship.ds <- DS.prior(ship, max.m = 4, g.par = c(.5,.5),
89               family = "Binomial")
90 surg.ds <- DS.prior(surg, max.m = 8, g.par = g.surg,
91               family = "Binomial", smooth.crit = "AIC")
92 # Generate plots
93 dev.new(height = 5, width = 15)
94 par(mfrow = c(1,3))
95 plot(rat.ds, plot.type = "mDev", main = "(a)")
96 plot(ship.ds, plot.type = "mDev", main = "(b)")
97 plot(surg.ds, plot.type = "mDev", main = "(c)")
98 ####
99 # EQUATIONS 2.2.16 – 2.2.19
100 ####
101 # Insurance DS
102 data(AutoIns)
103 g.ins <- gMLE.pg(AutoIns)
104 ins.ds <- DS.prior(AutoIns, max.m = 3, g.par = g.ins,
105               family = "Poisson")
106 # Display parameters
107 rat.ds
108 surg.ds
109 ship.ds
110 ins.ds
111 ####
112 # Figure 2.7: DS vs g plots
113 ####
114 dev.new( height = 10, width = 10)
115 par(mfrow = c(2,2))
116 plot(rat.ds, plot.type = "DSg", main = "(a)")

```

```

117 plot(surg.ds, plot.type = "DSg", main = "(b)")
118 plot(ship.ds, plot.type = "DSg", main = "(c)")
119 plot(ins.ds, plot.type = "DSg", main = "(d)", xlim = c(0,2))
120 ###
121 # Figure 2.8: MacroInference for rat, terb, and tacks
122 ###
123 # Rat, terb, and tacks macro
124 rat.macro <- DS.macro.inf(rat.ds, num.modes = 2,
125                           method = "mode", iters = 25)
126 terb.macro <- DS.macro.inf(terb.ds, method = "mean", iters = 25)
127 tack.macro <- DS.macro.inf(tack.ds, num.modes = 2,
128                           method = "mode", iters = 25)
129 # Generate plots
130 dev.new(height = 5, width = 12)
131 par(mfrow = c(1,3))
132 plot(rat.macro, main = "(a)")
133 plot(terb.macro, main = "(b)")
134 plot(tack.macro, main = "(c)")
135 ###
136 # Figure 2.9: Arsenic U-function and macro estimate
137 ###
138 # Arsenic DS
139 data(arsenic)
140 ars.start <- gMLE.nn(arsenic$y, arsenic$sse)$estimate
141 ars.ds.me <- DS.prior(arsenic, max.m = 4, g.par = ars.start,
142                    family = "Normal", LP.type = "MaxEnt")
143 # Arsenic macro
144 ars.mac.me <- DS.macro.inf(ars.ds.me, num.modes = 2,
145                          iters = 10, method = "mode")
146 # Generate plots
147 dev.new(height = 5, width = 8)
148 par(mfrow = c(1,2))
149 plot(ars.ds.me, plot.type = "Ufunc", main = "(a)")

```

```

150 plot(ars.mac.me, main = "(b)")
151 lines(ars.ds.me$prior.fit$theta.vals, ars.ds.me$prior.fit$parm,
152       lty = "dashed", col = "blue")
153 ###
154 # Figure 2.10: Rat elastic-Bayes (mean and mode)
155 ###
156 # Rat micro-reduce
157 rat.micro <- DS.posterior.reduce(rat.ds)
158 # Generate plots
159 dev.new(height = 5, width = 10)
160 par(mfrow = c(1,2))
161 plot(rat.micro$PEBMN, rat.micro$DSMN, main = "(a)")
162 abline(a=0, b = 1, col = "gray")
163 plot(rat.micro$PEBMD, rat.micro$DSMD, main = "(b)")
164 abline(a=0, b = 1, col = "gray")
165 ###
166 # Figure 2.13: Microinference
167 ###
168 # Surgical Node Micro
169 surg.micro.y1 <- DS.micro.inf(surg.ds, y.0 = 7, n.0 = 32)
170 surg.micro.y2 <- DS.micro.inf(surg.ds, y.0 = 3, n.0 = 6)
171 surg.micro.y3 <- DS.micro.inf(surg.ds, y.0 = 17, n.0 = 18)
172 # Generate plots
173 plot(surg.micro.y3$post.fit$theta.vals, surg.micro.y3$post.fit$ds.pos,
174       type = "l", col = "red", lwd = 2,
175       xlab = "", ylab = "")
176 lines(surg.micro.y2$post.fit$theta.vals, surg.micro.y2$post.fit$ds.pos,
177       col = "tomato")
178 lines(surg.micro.y1$post.fit$theta.vals, surg.micro.y1$post.fit$ds.pos,
179       col = "orange")
180 # Rat micro and plot
181 rat.micro <- DS.micro.inf(rat.ds, y.0 = 4, n.0 = 14)
182 plot(rat.micro)

```

```

183 # Shipyard micro and plot
184 ship.micro <- DS.micro.inf(ship.ds, y.0 = 0, n.0 = 5)
185 plot(ship.micro)
186 # Tacks micro
187 tacks.micro.y1 <- DS.micro.inf(tack.ds, y.0 = 3, n.0 = 9)
188 tacks.micro.y2 <- DS.micro.inf(tack.ds, y.0 = 6, n.0 = 9)
189 tacks.micro.y3 <- DS.micro.inf(tack.ds, y.0 = 8, n.0 = 9)
190 # Generate plots
191 dev.new(height = 5, width = 15)
192 par(mfrow = c(1,3))
193 plot(tacks.micro.y1, main = "(a)")
194 plot(tacks.micro.y2, main = "(b)")
195 plot(tacks.micro.y3, main = "(c)")
196 ###
197 # Figure 2.14a: Butterfly
198 ###
199 # Butterfly DS
200 data(CorbBfly)
201 bfly.g <- c(0.104, 89.79)
202 bfly.ds <- DS.prior(CorbBfly, max.m = 4, bfly.g,
203                    family = "Poisson")
204 # Generate plot
205 plot(bfly.ds)
206 ###
207 # Figure 2.15: Norberg Insurance Data
208 ###
209 # Norberg DS
210 data(NorbergIns)
211 nor.g <- gMLE.pg(NorbergIns$deaths, NorbergIns$exposure/344)
212 nor.ds <- DS.prior(NorbergIns$deaths/(NorbergIns$exposure/344),
213                  max.m = 4, nor.g, family = "Poisson")
214 # Generate plots
215 dev.new(height = 5, width = 10)

```

```

216 par(mfrow = c(1,2))
217 plot(nor.ds, plot.type = "Ufunc")
218 plot(nor.ds, plot.type = "DSg")
219 # Norberg macro and plot
220 nor.macro <- DS.macro.inf(nor.ds, num.modes = 2,
221                          method = "mode", iters = 25)
222 plot(nor.macro)
223 # Norberg micro and plot
224 nor.micro.y1 <- DS.micro.inf(nor.ds, y.0 = 4, e.0 = 0.4500291)
225 nor.micro.y2 <- DS.micro.inf(nor.ds, y.0 = 2, e.0 = 0.2506105)
226 nor.micro.y3 <- DS.micro.inf(nor.ds, y.0 = 57, e.0 = 19.11462)
227 # Generate micro plots
228 dev.new(height = 5, width = 15)
229 par(mfrow = c(1,3))
230 plot(nor.micro.y1)
231 plot(nor.micro.y2)
232 plot(nor.micro.y3)
233 ###
234 # Figure 2.17 (a,b)
235 ###
236 # Ulcer macro
237 ulc.mac <- DS.macro.inf(ulc.ds, method = "mean", iters = 25)
238 # Generate plot
239 plot(ulc.mac)
240 # Child Illness DS
241 data(ChildIll)
242 thai.g <- gMLE.pg(ChildIll)
243 thai.ds <- DS.prior(ChildIll, max.m = 10, thai.g,
244                  family = "Poisson")
245 # Generate plot
246 plot(thai.ds)

```