

ABSTRACT

Graphical Methods in Prior Elicitation

Christopher Casement, Ph.D.

Chairperson: David J. Kahle, Ph.D.

Prior elicitation is the process of quantifying an expert's belief into the form of a probability distribution on a parameter(s) to be used in a Bayesian data analysis. Existing methods require experts to quantify their belief by specifying multiple numerical distribution summaries, which are then converted into the parameters of a given prior family. The resulting priors, however, may not accurately represent the expert's opinion, which in turn can undermine the accuracy of the analysis.

In this dissertation we propose two interactive graphical strategies for prior elicitation, along with web-based Shiny implementations for each, that do not rely on an expert's ability to reliably quantify their beliefs. Instead, the expert moves through a series of tests where they are tasked with selecting hypothetical future datasets they believe to be most likely from a collection of candidate datasets that are presented in graphical form. The algorithms then convert these selections into a prior distribution on the parameter(s) of interest. After discussing each elicitation method, we propose a variation on the Metropolis-Hastings algorithm that provides support for the underlying stochastic scheme in the second of the two elicitation strategies. We apply the methods to data models that are commonly employed in practice, such as the Bernoulli, Poisson, and Normal, though the methods can be more generally applied to other univariate data models.

Graphical Methods in Prior Elicitation

by

Christopher Casement, B.A., M.A., M.S.

A Dissertation

Approved by the Department of Statistical Science

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

David J. Kahle, Ph.D., Chairperson

John W. Seaman, Jr., Ph.D.

James D. Stamey, Ph.D.

Dean M. Young, Ph.D.

John F. Tripp, Ph.D.

Accepted by the Graduate School
August 2017

J. Larry Lyon, Ph.D., Dean

TABLE OF CONTENTS

LIST OF FIGURES	vi
LIST OF TABLES	viii
ACKNOWLEDGMENTS	ix
DEDICATION	x
1 Introduction	1
2 Deterministic Procedures for Graphical Prior Elicitation in Univariate Models	3
2.1 Introduction	3
2.2 Prior Elicitation	5
2.2.1 Current Methods and Tools	5
2.2.2 Drawbacks to Current Methods	7
2.3 Graphical Inference Procedures	8
2.3.1 Rorschach Procedure	9
2.3.2 Line-up Procedure	11
2.4 Graphical Elicitation	13
2.4.1 Data Models with One Unknown Parameter	13
2.4.2 $\mathcal{N}(\mu, \sigma^2)$ Data Model with μ and σ^2 Unknown	21
2.5 Implementation	31
2.6 Conclusion	33
3 Stochastic Procedures for Graphical Prior Elicitation in Univariate Models	35
3.1 Introduction	35

3.2	Prior Elicitation	37
3.2.1	Existing Methods and Implementations	37
3.2.2	Shortcomings of Existing Methods	38
3.3	Graphical Procedures for Statistical Inference	39
3.3.1	Rorschach Procedure	40
3.3.2	Line-up Procedure	41
3.4	Graphical Elicitation	43
3.4.1	Data Models with One Unknown Parameter	44
3.4.2	$\mathcal{N}(\mu, \sigma^2)$ Data Model with μ and σ^2 Unknown	48
3.5	Shiny Application	51
3.6	Technical Considerations	53
3.7	Conclusion	56
4	The Rigid Metropolis Algorithm for Approximate Bayesian Computation	57
4.1	Introduction	57
4.2	Total Variation Distance	59
4.3	Rigid MCMCs	61
4.3.1	General Scheme	61
4.3.2	Applications to Graphical Prior Elicitation	62
4.4	Conclusion	78
5	Conclusion	79
A	Effective Sample Size Computations	82
A.1	Data Model and Prior Combinations with Closed Form ESS Expressions	82
A.2	$\mathcal{N}(\mu, \sigma^2)$ Data Model with μ and σ^2 Unknown and a Normal-Inverse-Gamma Joint Prior	82
	BIBLIOGRAPHY	88

LIST OF FIGURES

2.1	Sixteen histograms, each of 100 random samples from a $\mathcal{N}(0, 1)$ distribution.	10
2.2	Sixteen scatterplots: fifteen of simulated data, one of real data randomly mixed in.	12
2.3	Algorithm steps for a Bernoulli data model. Check marks indicate the expert's selection.	20
2.4	Algorithm steps for a $\mathcal{N}(\mu, \sigma^2)$ data model with μ and σ^2 unknown. Check marks indicate the expert's selection.	23
2.5	Two-dimensional view of the algorithm steps for a $\mathcal{N}(\mu, \sigma^2)$ data model with μ and σ^2 unknown. Numbers in red indicate the expert's selection.	24
2.6	Expert inputs to the Shiny app for graphical elicitation of the Bernoulli p	31
2.7	Rorschach training plots.	32
2.8	Example of graphs presented to the expert at each step.	32
2.9	Elicited prior and its summaries.	33
2.10	History plot of the proportions at each step.	34
3.1	The MATCH Uncertainty Elicitation Tool is a free browser-based JavaScript elicitation tool.	38
3.2	BetaBuster is a free Java application used for mode-percentile elicitation of beta distributions.	39
3.3	Sixteen histograms, each of 500 random observations from the Beta $(0.5, 0.5)$ distribution.	41
3.4	Fifteen scatterplots of simulated residuals, one of real residuals randomly mixed in.	43
3.5	Before the graphical selection process begins, the expert can go through Rorschach training. Each bar chart displays a random sample of size $n = 100$ from the Bernoulli($p = 0.67$) distribution.	51
3.6	At each step of the selection process two graphics are presented to the expert, along with options for choosing between them.	52

3.7	After the expert finishes the selection process, information regarding the elicited prior is presented.	53
3.8	A trace plot of the accepted proportions is also available.	53
4.1	Average total variation distances for Bernoulli(p) data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, 1\}$. Plot labels communicate Bernoulli probabilities used.	70
4.2	Average total variation distances for Poisson(λ) data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, 1\}$. Plot labels communicate Poisson rates used.	71
4.3	Average total variation distances for $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, 1\}$. Plot labels communicate Normal means used.	71
4.4	Average total variation distances for Bernoulli(p) data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, p_3^{\text{rigid}}, 1\}$. Plot labels communicate Bernoulli probabilities used.	73
4.5	Average total variation distances for Poisson(λ) data models based on rigid acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, p_3^{\text{rigid}}, 1\}$. Plot labels communicate Poisson rates used.	74
4.6	Average total variation distance for $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, p_3^{\text{rigid}}, 1\}$. Plot labels communicate Normal means used.	74
4.7	Histograms of 1,000 total variation distances for Bernoulli(p) data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$. Plot labels communicate Bernoulli probabilities used.	76
4.8	Histograms of 1,000 total variation distances for Poisson(λ) data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$. Plot labels communicate Poisson rates used.	77
4.9	Histograms of 1,000 total variation distances for $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$. Plot labels communicate Normal means used.	77

LIST OF TABLES

4.1	Total variation distance summaries for Bernoulli(p), Poisson(λ), and $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on $\mathcal{P} = \{10^{-6}, 1\}$, indicating that merely rounding the transition probability in the Metropolis algorithm yields an unacceptably poor rigid approximation to standard Metropolis.	69
4.2	Total variation distance summaries for Bernoulli(p), Poisson(λ), and $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on $\mathcal{P} = \{10^{-6}, p_{2*}^{\text{rigid}}, 1\}$, indicating that three rigid probabilities improve significantly on two (simple rounding), yet still leaving around a 10% worst-case error.	72
4.3	Total variation distance summaries for Bernoulli(p), Poisson(λ), and $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$, indicating that four rigid probabilities improve on three, leaving a worst-case error on the order of 5%.	75
A.1	Effective sample size expressions for three common cases.	82

ACKNOWLEDGMENTS

Thank you Dr. Kahle for your dedication to making me a better scholar and educator, and for your constant encouragement throughout my studies at Baylor. I couldn't have asked for a better mentor. Thank you Dr. Stamey and Dr. Tubbs for helping me achieve my dream of teaching in higher education, and for creating an environment at Baylor that made me feel at home. Thank you to my committee members and the graduate faculty for all you've done for me and my peers. Thank you also to my Baylor friends for your friendship. And most of all, thank you to my family – I love you all and couldn't have done any of this without you.

DEDICATION

To my amazing parents and grandparents, for your unwavering love and support.

CHAPTER ONE

Introduction

Prior specification is fundamental to the Bayesian paradigm. Informative priors can allow analysts to inject expert opinion directly into the modeling process, but since they can strongly influence an analysis, using them demands a principled approach to prior specification. Prior elicitation refers to this process of quantifying an expert's belief into a probability distribution. Existing elicitation schemes ask experts to quantify their belief by providing two or more distribution summaries, which are then converted into the parameters of a given family.

While software exist to assist experts in the process, difficulties remain that can lead to a prior distribution that does not accurately reflect expert opinion. To this end, we propose two interactive graphical strategies for prior elicitation, along with web-based Shiny implementations for each, that do not rely on quantities experts may not be able to reliably attest to. Instead, the expert moves through a series of tests where they are tasked with selecting hypothetical future datasets they believe to be most likely from a collection of candidate datasets that are presented in graphical form. The algorithms then convert these selections into a prior on the parameter(s) of interest. We apply the methods to data models that are commonly employed in practice, such as the Bernoulli, Poisson, and Normal, though the methods can be more generally applied to other univariate data models.

In Chapter 2 we present the first of the two graphical elicitation methods, which is deterministic in its underlying structure. The expert provides the number of times they have seen a particular scenario before – a prior effective sample size – and then moves through the graphical selection process. At each step they are presented five graphics and must select the one they believe is most likely as a future hypothetical

dataset. The hyperparameters for the chosen prior family are then found based on a combination of the expert's selections and the prior effective sample size.

In Chapter 3 we present the second graphical elicitation method, which is built upon a stochastic foundation. At each step the expert is presented with two graphics and five options representing how likely the graphics are relative to one another. These options are based on optimized values computed using rigid Metropolis, a variation on the Metropolis sampler we discuss in Chapter 4. After making a number of selections, the MLE of the chain that results from the selections is computed, leading to the hyperparameters of the specified prior family.

We next discuss rigid Metropolis in Chapter 4. These procedures provide support for the underlying stochastic scheme proposed in Chapter 3. Rigid Metropolis is a variation on the standard Metropolis algorithm that allows for discrete sets of transition probabilities and which helps determine the expert's options in the stochastic graphical procedure. We lastly conclude with a summary of the dissertation and ideas for future work in Chapter 5.

CHAPTER TWO

Deterministic Procedures for Graphical Prior Elicitation in Univariate Models

Abstract

Standard prior elicitation procedures require experts to explicitly quantify their beliefs about parameters in the form of multiple summaries. In this chapter we draw on recent advances in the statistical graphics and information visualization communities to propose a novel elicitation scheme that implicitly learns an expert's opinions through their sequential selection of graphics of carefully constructed hypothetical future samples. While the scheme can be applied to a broad array of models, we use it to construct procedures for elicitation in data models commonly used in practice: Bernoulli, Poisson, and Normal. We also provide open-source, web-based Shiny implementations of the procedures.

2.1 Introduction

The use of prior distributions is an essential part of the Bayesian framework. Informative priors allow expert belief to be included in statistical analyses; however, special care must be taken when selecting such priors, since they can strongly and inappropriately impact conclusions by overwhelming the data. The process of quantifying expert opinion into a prior distribution is known as prior elicitation.

Garthwaite et al. (2005) provides a nice introduction to the prior elicitation process and summarizes the existing literature. They divide the elicitation process into four stages: setup, elicit, fitting, and adequacy check. In the setup stage, the statistician (or a facilitator) explains the elicitation process to the expert and trains him or her in the probabilistic concepts required for the exercise, emphasizing aspects of uncertainty and variability that may not be familiar to the expert. In the elicit stage,

the expert provides quantitative summaries of their beliefs, such as a mean, mode, or percentile. In the fitting stage, the statistician converts the expert’s summaries into a probability distribution, typically in the form of the standard parameters of a given prior family. In the adequacy check stage, the statistician communicates to the expert the implications of their summaries in order to make sure they correspond to the expert’s actual beliefs. This last step is considered essential, since errors can creep in at many places along the way and undermine the accuracy of the analysis.

Statistical discussions concerning elicitation typically center around the elicit and fitting stages: what quantities should be elicited, how they can be converted into the standard parameters of a given prior family, what the properties of the conversion process are, etc. While these discussions are important, the resulting methods often undervalue the importance of the entire process. Most of the methods can be done with no concrete training of the expert and without assessing the validity of the resulting prior.

Inspired by recent advances in the statistical graphics and information visualization communities, in this chapter we propose a graphical scheme for prior elicitation that addresses each of the stages of the prior elicitation process in a holistic manner. To properly train the expert in the kinds of uncertainty that can be expected, the proposed scheme uses the Rorschach procedure (Buja et al., 2009), which involves drawing multiple sets of pseudo-random variates from the same data model and then plotting them side-by-side for the expert to examine. In the elicit stage, unlike standard elicitation methods that require the expert to explicitly specify values that can be difficult to quantify, the proposed scheme implicitly learns an expert’s opinions through their sequential selection of graphics of carefully constructed hypothetical future samples. The selections result in summaries of the prior that are easily converted into the relevant hyperparameters in the fitting stage. Since the specifications are made through graphics of hypothetical future samples, the adequacy check is

actually a part of the elicit stage; however, the implementation allows for further exploration of the elicited prior as well.

While the scheme is conceptually quite general, we use it to develop specific, detailed procedures to elicit priors for univariate data models: a beta prior on the Bernoulli proportion p , a gamma prior on the Poisson rate λ , a Normal prior on the Normal mean, μ , and a Normal-inverse-gamma prior on the Normal mean and variance, $[\mu, \sigma^2]'$, all very important cases in practice. To enable others to use the procedures, we offer open-source, web-based Shiny implementations of all four and use the population proportion implementation to illustrate the scheme. The free application is available online at `ccasement.shinyapps.io/graphicalElicitation`.

The chapter proceeds as follows. In Section 2.2 we examine the prior elicitation process, including current strategies and their drawbacks. In Section 2.3 we discuss two recent advances in the statistical graphics literature, the Rorschach and line-up procedures, that lay the conceptual foundation for the proposed graphical elicitation scheme. Section 2.4 then formally presents the proposed method. In Section 2.5 we describe the free Shiny implementation and show an example of the proposed process. We then conclude with a summary of the chapter and ideas for further research in Section 2.6.

2.2 Prior Elicitation

2.2.1 Current Methods and Tools

Many methods have been proposed for prior elicitation. Of the existing methods, virtually all are analytical in nature, requiring the expert to quantify their beliefs concerning the parameter of interest and the statistician to convert the specifications into the standard parameters of the family. When possible, the summaries elicited are on an observational scale: they are on a scale that is familiar to the expert.

A simple example describes the general flavor of the methods: suppose a beta prior on a population proportion is desired. Clearly, the statistician cannot simply ask a domain expert which α and β represent their belief, so he/she asks them about quantities that, if known, could be translated into α and β . Preferably, the quantities are values that the expert can reliably attest to. To that end, the mode and percentile (MP) method asks the expert two questions: 1) what is the most likely value of the proportion?,¹ and 2) what is the smallest (or largest) value that the proportion could possibly be? These are then taken to be the mode and a percentile of a Beta(α, β) distribution, and an inversion process is used to determine which α and β are consistent with the specifications (Wu et al., 2008).

Variants of the above strategy are standard practice in the elicitation community. These include the cumulative distribution function (CDF) method, the probability density function (PDF) method, and the equivalent prior sample (EPS) method, among others. For more in-depth discussions of these methods of prior elicitation, see Garthwaite et al. (2005), O’Hagan et al. (2006), and Kahle et al. (2014).

In addition to the methods, various software exist to assist users in performing the necessary computations. One tool, SHELF (the Sheffield Elicitation Framework), was developed at the University of Sheffield and is a package of documents, templates, and software that run in R (Oakley and O’Hagan, 2010). It goes well beyond mere computations, providing state-of-the-art descriptions of each part of the elicitation process, from the interaction with the expert to software implementations of many methods. A related tool is the MATCH (the Multidisciplinary Assessment of Technology Centre for Healthcare) Uncertainty Elicitation Tool. Produced by a team including some of the developers of SHELF, MATCH is based on SHELF but offers certain advantages over the package-based software (Morris et al., 2014); in particular, MATCH runs through a web browser via a JavaScript implementation, is fast,

¹ To make this question more observational, one might ask something akin to “Out of 100 patients, how many would you expect to present with ailment x ?”

and allows for session saving and sharing.² BetaBuster is another elicitation tool; it implements the MP method described above in Java to obtain a beta prior for a Bernoulli proportion and distributes it online gratis (Su, 2006). Another tool for prior elicitation, which runs through Microsoft Excel, enables its users to visualize the impact of changing parameter values on the shapes of distributions through the use of sliders (Jones and Johnson, 2014). As some of the details in the conversion process can be quite technical (specifically during the fitting stage), these implementations are indispensable contributions to the community of Bayesian statisticians.

2.2.2 Drawbacks to Current Methods

While the analytical methods discussed previously allow analysts to elicit priors, they nevertheless suffer from a number of drawbacks. One concern is that an expert may not be able to accurately quantify their beliefs in the form of statistical summaries, regardless of the extent of their specialized knowledge. The sensitivity of the MP method, for example, varies across the specification space, but in the wrong place even minor rounding misspecifications, say 2% changes in the percentile value, can result in priors that are significantly more informative (Blair, 2017).

Another drawback to analytical methods is that they are decoupled from the adequacy check stage of the elicitation process, which makes them error prone. As noted in Section 2.2, an essential part of any prior elicitation exercise is a post-elicitation assessment of the implications of the results on an observational scale. For example, in the MP method scenario described above, it is essential that once a prior has been elicited, data consistent with that prior be sampled from the data model and shown to the expert. If the analyst is not diligent in their duties and does not perform such reviews with the expert, an inadvertently unrepresentative prior may result with the potential to overwhelm a collection of data and undermine otherwise

² The MATCH elicitation tool is available at optics.eee.nottingham.ac.uk/match/uncertainty.php

valid conclusions. In fact, studies have shown that individuals, including experts, tend to be overconfident in their beliefs, often overestimating how precise their estimates are (Van den Steen, 2011 and references therein). Thus, the decoupling of analytical methods from the review stage predisposes experts to potentially very costly mistakes. This problem is shared, to varying degrees, by all analytical methods.

Yet another drawback to analytical methods is that the conversion process is not always cleanly possible: there is not always a bijection between the set of all possible values for the provided summaries, the “elicitation space,” and the set of all values for the prior parameters of a chosen family, the canonical parameter space, and moreover even when such a bijection exists it is not always simple – closed form expressions for the conversions need not exist and numerical routines may be unstable. Situations can arise where there is either no distribution from the chosen prior family that satisfies the summaries provided by the expert, or where there are multiple distributions in the prior family that satisfy the summaries. For instance, using the MP method for the elicitation of a population proportion, it is possible that no beta distribution satisfies the mode and percentile specifications, or exactly two do (Wu et al., 2008). In that same scenario, the required conversions do not have simple closed form expressions.

These drawbacks suggest the search for a different paradigm for prior elicitation – one that moves away from quantitative statistical summaries and yet maintains and maximizes the interaction of the expert with the observational scale of the data. The graphical scheme described in Section 2.4 satisfies both of these criteria.

2.3 Graphical Inference Procedures

We now introduce two statistically rigorous procedures, the Rorschach and line-up procedures (Buja et al., 2009, Wickham et al., 2010), that operate in the realm of graphical hypothesis testing and motivate the proposed graphical elicitation method.

The goal of the Rorschach is to help users better understand natural variability in data and grow more accustomed to working with it. The line-up is a kind of graphical goodness-of-fit test. In this section we briefly present both of these procedures before leveraging them in Section 2.4 for prior elicitation.

2.3.1 *Rorschach Procedure*

Considering the expert as an instrument, the purpose of the Rorschach procedure is to help calibrate them to natural variability in data. For the procedure G datasets are simulated from the same data distribution and an appropriate graphic (e.g. histogram) is plotted for each. The graphics are then placed side-by-side for visual inspection and comparison. The individual performing the procedure examines the graphics and looks for patterns and deviations from what they would expect. Since the data are all drawn from the same distribution, often called the “null” distribution, apparent oddities observed in the graphics are actually the result of natural variability in the data. More technically, the underlying null distribution induces a probability distribution on the space of graphics that is learned by the expert through the Rorschach procedure.

For example, suppose we want to train the expert in the natural variability of the $\mathcal{N}(0, 1)$ distribution with 100 observations as seen in histograms. The Rorschach procedure suggests that we generate $G = 16$ datasets (say), each of size 100, from the standard normal distribution and then construct a histogram for each using the same axes scales. The Rorschach procedure is illustrated in Figure 2.1, with the number of bins in each histogram determined by Scott’s rule (Scott, 1979).

The difference between what may be expected – a histogram with a nice clean bell shape – and what is observed illustrates the natural variability associated with sampling from the $\mathcal{N}(0, 1)$, or more generally the null, distribution. The purpose of the Rorschach procedure is for the expert to examine and learn what sorts of

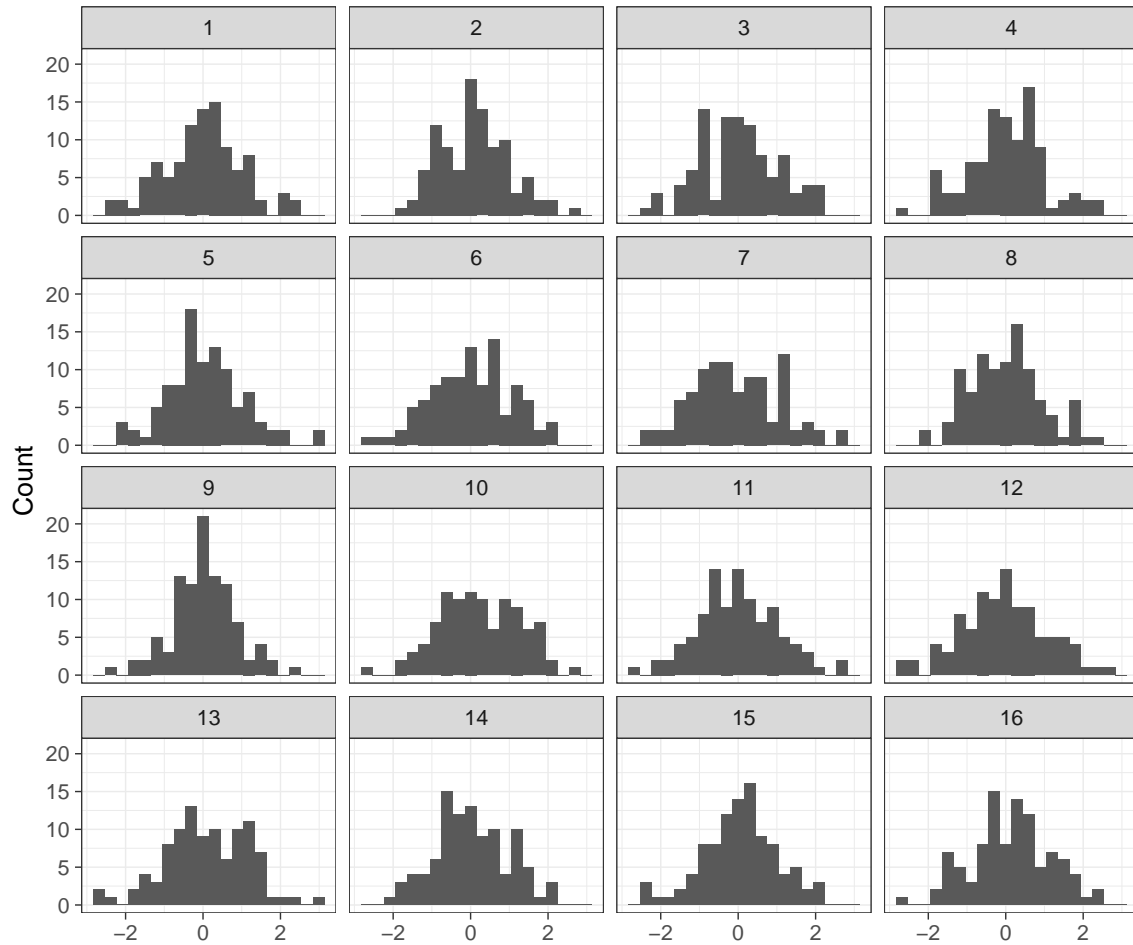


Figure 2.1. Sixteen histograms, each of 100 random samples from a $\mathcal{N}(0, 1)$ distribution.

deviations from the null distribution can be expected and consequently be less likely to reject the notion that a dataset seen through a graphic is inconsistent with a particular distribution merely because the graphic does not illustrate features expected in the ideal case. In this sense the procedure prepares experts for a kind of visual goodness-of-fit test. In fact, a simple regeneration of the graphics allows them to run the procedure several times.

2.3.2 Line-up Procedure

The line-up procedure, another method in the realm of visual hypothesis testing, is the graphical goodness-of-fit test that naturally arises from the Rorschach procedure. Like other goodness-of-fit tests, the null hypothesis is that the data come from a specific distribution, and the alternative is that the data do not come from that distribution. When testing a model, the model is fit and the estimated distribution is used as the null. $G - 1$ datasets, each of the same size as the original dataset, are then simulated from the null (Wickham et al., 2010), and appropriate graphs are plotted side-by-side. The series of simulations can be thought of as $G - 1$ parametric bootstrap replications of the original dataset.

Unlike the Rorschach procedure, with the line-up the graph of the real data is randomly mixed into the $G - 1$ replicate graphics. The user is then tasked with selecting the graph of the real data from the set of G plots. If they successfully do so, they are believed to have provided sufficient evidence to reject the null hypothesis, suggesting the true data do not come from the null distribution. In other words, the model does not counterfeit key features the expert expects to see in the data.

As an example, consider the simple regression problem posed by a dataset of size 50 that consists of one continuous response and one continuous explanatory variable. A simple linear model can be tested as follows: fit the model with the data, generate 15 synthetic datasets of size 50 from the fitted model, and create a scatterplot of each. Then array the 15 scatterplots side-by-side, with the scatterplot of the original data, a sixteenth graphic, randomly mixed in. The user is then tasked with selecting which of the 16 scatterplots corresponds to the actual data. If they are successful at selecting the real data, the null hypothesis of the simple linear model is rejected. We illustrate such a procedure in Figure 2.2 (for the solution to this example, see footnote ³).

³ The real data are displayed in graph 9. This scatterplot exhibits a nonlinear trend, while the other 15 graphs are of data with linear trends.

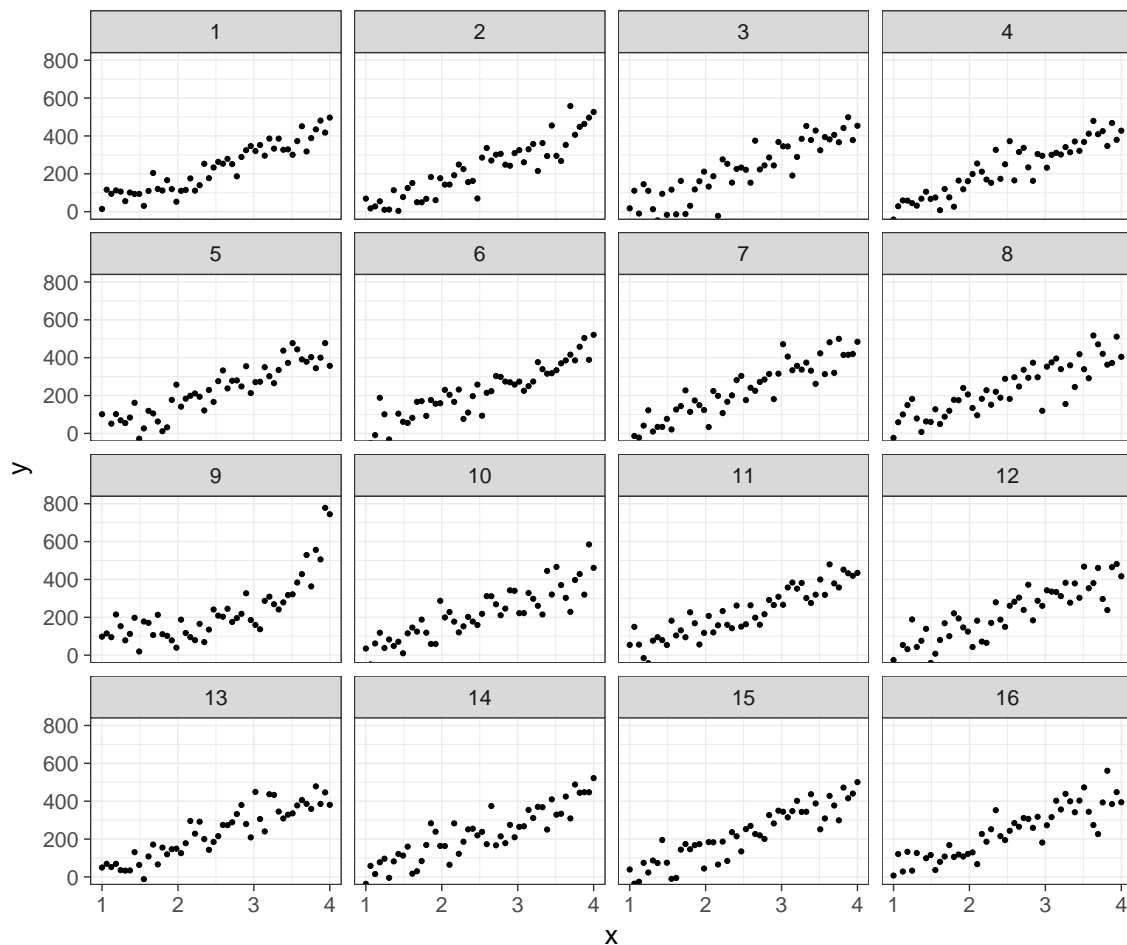


Figure 2.2. Sixteen scatterplots: fifteen of simulated data, one of real data randomly mixed in.

There are many variants of the procedure that might be and have been proposed; we have merely included the basic concepts to prepare the reader for the graphical elicitation strategies described next. There are two key advantages of the visual procedure (the line-up) over conventional methods. First, since it is rooted in basic techniques of exploratory data analysis, it is easily understood by non-statisticians. Indeed, it has also been proposed as a statistics education tool to understand goodness-of-fit testing for just this reason (Buja et al., 2009). Second, the strategy depends on the user's sense of meaningful difference. Human cognition determines significance, both practical and statistical. Both of these considerations seem ideally tailored to the

prior elicitation purposes we describe in the next sections. We refer the reader to the articles above to provide a more complete picture of the Rorschach and line-up procedures as well as for more examples.

2.4 Graphical Elicitation

We now introduce the interactive graphical scheme that is the main contribution of this work. The scheme not only assists users in eliciting prior distributions but also addresses the limitations of the existing methods discussed in Section 2.2, namely (1) the difficulty of experts to accurately quantify their beliefs, (2) the decoupling of the elicitation stage and the assessment stage, and (3) the mathematical challenges presented by the conversion between the elicitation space and the standard parameter space of the desired prior family. The proposed method, which is deterministic in its underlying structure, involves the expert sequentially making line-up-like judgments on data graphics, which are then processed to determine a prior. After a description of the general scheme and the intuition that motivates it, we describe specific procedures for eliciting priors on the Bernoulli p , the Poisson λ , and the Normal μ (σ^2 known) in Section 2.4.1, followed by the Normal $[\mu, \sigma^2]$ (both unknown) in Section 2.4.2, although the same logic can be more generally applied to other models.

2.4.1 Data Models with One Unknown Parameter

2.4.1.1 Description of the general scheme. Suppose a data model $\mathcal{M} = \{f_\theta\}_{\theta \in \Theta}$ is assumed with $\Theta \subset \mathbb{R}$ and a prior p_η is desired for θ from the family $\mathcal{P} = \{p_\eta\}_{\eta \in \mathbf{H}}$. We assume that the hyperparameter space \mathbf{H} is two-dimensional, as it is in the cases of interest in this dissertation; however, this can likely be relaxed with modifications. Since \mathbf{H} is two-dimensional, at least two quantities need to be elicited from the expert. In analytical terms, the quantities obtained in the proposed graphical method are the mode of the prior and its prior equivalent sample size (ESS). The mode is

the θ the expert believes most likely, and the ESS can be thought of as a measure of the expert’s uncertainty. The ESS is a way of looking at a prior as the posterior of a previously observed dataset with a reference prior; it is the number of observations required to obtain the prior as a posterior (Christensen et al., 2011, Gelman et al., 2014). In many cases of practical importance, including the applications here, it is a simple function of the parameters of the distribution, but in general it does not have a closed form and must be computed through simulation (Morita et al., 2008).

If the model is correct, the data generating mechanism is f_{θ^*} for some unknown $\theta^* \in \Theta$. At the heart of the graphical procedure is the determination of the mode of the prior, the value the expert believes θ^* most likely to be, and so to understand how the procedure works it is easiest to begin there. The general intuition is this: obtain a reasonable range $[l, u]$ for θ^* and then hone in on its value through a series of graphics generated with values from increasingly narrow ranges. Then, combine the information with a measure of uncertainty (the ESS) to arrive at the expert’s prior, represented in terms of $\boldsymbol{\eta}$, not the mode and ESS. We now describe the process in more detail.

The first step of the scheme is to determine the reasonable range $[l, u] \subset \Theta$ that contains θ^* . We assume that Θ is an open interval of the real line. Determining $[l, u]$ is easy in cases where Θ is bounded, since we can simply select values near the boundary $\partial\Theta$. In cases where Θ is unbounded, for the unbounded side(s) we ask the expert for the smallest (largest) value an outcome could possibly be, denoted, x_l (x_u), and select l (u) as the lowest (highest) θ for which x_l (x_u) is the 1st (99th) percentile. With the reasonable range $[l, u]$ in hand, we construct a G -point equispaced mesh on $[l, u]$ and generate N pseudo-random variates from f_{θ} , using each of the G mesh values as θ . We construct graphics for those G scenarios, permute them, and present them to the expert, asking him/her to select the one that would be most likely as a future

dataset of N observations⁴. Once a selection is made, the underlying parameter used to generate the graphic is recorded, and a slightly narrower range of parameter values is constructed around the new parameter value. Again G values in the range are taken, pseudo-random variates are generated, plotted, and permuted, and the expert is asked for their opinion. This process is repeated a number of times until the interval is a suitably small length, at which point the algorithm declares convergence, a result guaranteed by the shrinking interval width at each iteration. The average of the last K selections, denoted $\bar{\theta}_K^*$, is taken to be the value the expert believes θ^* most likely to be, the mode of their prior. The average is simply used to decrease the variability of the resulting estimate.

While the intuition behind the graphical selections was described first, the determination of the ESS and the Rorschach training actually take place before the expert’s graphical selections. At the beginning of the exercise the expert is asked for the number of observations his/her experience is based on, denoted n . The expert is then asked for a typical observation and lead through a Rorschach training procedure; see Section 2.5 for details. After training, the expert is presented with the graphical selection process described above. Setting $\text{ESS} = n$,⁵ and the mode equal to $\bar{\theta}_K^*$, then yields the nonlinear system

$$\text{ess}(p_{\boldsymbol{\eta}}) = n \quad (2.1)$$

$$\text{mode}(p_{\boldsymbol{\eta}}) = \frac{1}{K} \sum_{j=S-(K-1)}^S \theta_j^* \stackrel{\text{def}}{=} \bar{\theta}_K^*, \quad (2.2)$$

where $\{\theta_j^*\}_{j=1}^S$ are the S parameter values corresponding to the graphical selections made by the expert. A solution $\boldsymbol{\eta}^*$ to this system is the vector of hyperparameters for the expert’s prior. The full process is detailed in algorithmic format in Algorithm 1.

⁴ Choosing one graphic may prove challenging if, for instance, the expert believes multiple graphics are equally likely for a hypothetical future sample. He/she is still tasked with choosing one, though.

⁵ We set $\text{ESS} = n$ instead of $\text{ESS} = N$ so that the ESS is tied to the expert’s past experiences.

input : n – the number of previous observations
 N – the number of observations in a hypothetical future sample
 f_θ – the data model family parameterized by $\theta \in \Theta$
 p_η – the prior model family parameterized by $\eta \in \mathbf{H}$
 r – the rate at which the window diminishes on the link scale
(e.g. 15%)
 g – a link function $g : \Theta \rightarrow \mathbb{R}$ (typically the canonical link:
 $g(p) = \text{logit}(p)$ for the binomial; $g(\lambda) = \log \lambda$ for the Poisson)
 G – the mesh size/number of graphics to be presented (e.g. 5)
 K – the number of final selections to average for $\bar{\theta}_K^*$ (e.g. 5)
 tol – a tolerance for declaring convergence (e.g. 2%)

output: η^* , the hyperparameters for the prior on Θ

```

1 Construct an interval  $[l, u] \subset \Theta$ 
2  $w \leftarrow g(u) - g(l)$ 
3  $j \leftarrow 1$ 
4 repeat
5   for  $i \leftarrow 1$  to  $G$  do
6      $\theta_i \leftarrow l + (i - 1) \frac{u-l}{G-1}$ 
7     Sample  $\mathbf{d}^{(i)} = [d_1^{(i)} \dots d_N^{(i)}] \stackrel{iid}{\sim} f_{\theta_i}$ 
8   end
9   Construct faceted (trellised) graphics with  $\mathbf{d}^{(1)}, \dots, \mathbf{d}^{(G)}$ 
10   $\theta_j^* \leftarrow$  the  $\theta_i$  corresponding to the plot of the expert's selection
11   $w \leftarrow (1 - r)w$ 
12   $(l, u) \leftarrow g^{-1} \left( g(\theta_j^*) \pm \frac{w}{2} \right)$ 
13   $j \leftarrow j + 1$ 
14 until  $u - l < tol$ ;
15  $\eta^* \leftarrow$  Solution to Equations (2.1) and (2.2) for  $\eta$ 

```

Algorithm 1. General scheme of the procedure for data models with one unknown parameter.

2.4.1.2 Implementation notes. A few implementation details of Algorithm 1 warrant further discussion.

First, after the expert makes a selection, the underlying parameter is transformed using a link function – logit for a Bernoulli model, log for a Poisson model, and identity for a Normal model (variance known). Once endpoints for a new interval are determined on the link scale, they are converted back to the parameter scale using

the appropriate inverse-link function. Such a transformation ensures all values in the new interval fall within the appropriate support Θ : $(0, 1)$ for proportions and $(0, \infty)$ for rates. As the selection process proceeds, working on the link scale also allows the ranges to go beyond the initial $[l, u]$.

Second, the interval width reduction for the determination of the new interval (Step 11 of Algorithm 1) may need to be adjusted slightly for some cases (e.g. the Poisson) to guarantee a reasonably fast narrowing of the intervals from one step to the next. In the implementations described in Section 2.5, each interval (on the link scale) is 85% of the length of the previous interval for the Bernoulli and Normal cases, and 90% for the Poisson case. These values achieve a reasonable convergence rate for the algorithm: if the algorithm converges too slowly it requires many selections and expert fatigue might arise, whereas if the algorithm converges too quickly, natural variability in the data model may pigeonhole the expert into a region of the parameter space they don't believe to be likely, rendering inaccurate and unreliable results.

Last, we address the proper selection of N , G , and K . Recall that N is the sample size of the hypothetical future sample. For the number of random samples generated at each step, $N = 100$ is not so small that very unrepresentative samples are likely to be drawn. For instance, for a Bernoulli(0.5) data model, obtaining zero successes out of $N = 10$ random observations is more likely to occur than for $N = 100$ observations. Also, if N is too large and the data model displays features not fully consistent with the expert's beliefs, he/she may begin to question the accuracy of the process; $N = 100$ is not so large that such a problem will arise often. Recall G is the number of graphics presented at each step. We find that $G = 5$ allows for sufficient variability among the plots while limiting the potential fatigue and difficulty associated with having to compare many at once. Recall that K is the number of final selections averaged for the mode of the prior. While K depends on the convergence rate of the algorithm and the chosen tolerance, we find $K = 5$ to be suitable for a

broad array of circumstances. It also strikes a reasonable balance between taking only the final selection ($K = 1$), which potentially can result in a slightly biased mode due to natural variability in the final selection, and the average of all selections ($K = S$), which can be biased by initial guesses. The choice of K is thus important for addressing this bias-variance tradeoff. In general, a small K is preferable to a large K since the width of the interval is decreasing.

2.4.1.3 More on the prior inversion process. We now discuss the conversion of $\bar{\theta}_K^*$ and n to $\boldsymbol{\eta}^*$ for the three cases of interest.

A unique mode only exists for distributions in the Beta(α, β) family if $\alpha, \beta > 1$ (recall that Beta(1, 1) is the uniform distribution). Thus, the mode/ESS specification is a slight reduction in the parameter space, but is otherwise not very consequential. The ESS of a Beta(α, β) distribution is equal to $\alpha + \beta$. If the mode and ESS are specified, solving Equations (2.1) and (2.2) can be done with a simple algebraic inversion:

$$\alpha = \text{mode}(\text{ess} - 2) + 1 \quad (2.3)$$

$$\beta = \text{ess} - \text{mode}(\text{ess} - 2) - 1. \quad (2.4)$$

There are many similarities between the beta prior for the Bernoulli data model and the gamma prior for the Poisson data model. In the Poisson case, a unique mode only exists for distributions in the Gamma(α, β) family⁶ if $\alpha \geq 1$, which is an inconsequential reduction in the parameter space. The ESS is also simple: the ESS of a Gamma(α, β) is β . And, as in the beta case, a simple algebraic reformulation shows that the parameter spaces are in bijection:

$$\alpha = \text{mode}(\text{ess}) + 1$$

$$\beta = \text{ess}.$$

Unlike the Gamma and Beta families, the Normal family does not have any restrictions on the existence of the mode, guaranteeing a mode will always exist.

⁶ The parameterization used has mean α/β .

Similar to these other two families, however, the ESS is simple to calculate: the ESS of a $\mathcal{N}(\mu_0, \sigma_0^2)$ prior distribution on μ is equal to σ^2/σ_0^2 , where σ^2 is the known variance and σ_0^2 the prior variance of μ . Once again, the parameter spaces are in bijection:

$$\begin{aligned}\mu_0 &= mode \\ \sigma_0^2 &= \sigma^2/ess.\end{aligned}$$

2.4.1.4 Example: Bernoulli data model. We now present a concrete example using the Bernoulli model to clarify any lingering ambiguities.

Suppose the expert selects a Bernoulli data model and specifies he/she has seen a particular scenario of interest $n = 20$ times. A mesh of $G = 5$ equispaced proportions, $\{0.05, 0.275, 0.5, 0.725, 0.95\}$, is then created on the interval $[l, u] = [0.05, 0.95]$. These initial proportions are displayed on the first number line in Figure 2.3. Next, $N = 100$ random observations are drawn independently from Bernoulli distributions with these probabilities of success (100 from a Bernoulli(0.05) distribution, 100 from a Bernoulli(0.275) distribution, etc.). Bar charts for the five datasets are then displayed side-by-side, with their order randomly permuted to help combat selection bias that might emerge if the plots had a clear trend in their generating parameters. The facet labels shown in Figure 2.3 are not presented to the expert. The expert next selects the graph that appears most likely as a future dataset, and the proportion associated with the selected graph is stored.

The selected parameter value is then logit-transformed, and two new interval endpoints are calculated around the selected value such that the width of the resulting interval is 15% narrower than that from the previous step, $\text{logit}(0.95) - \text{logit}(0.05)$. These endpoints are then converted back to the $(0, 1)$ scale using the inverse-logit (logistic) function, and a new mesh of five equispaced values is created between the new endpoints. Five sets of 50 random samples are generated from Bernoulli distribu-

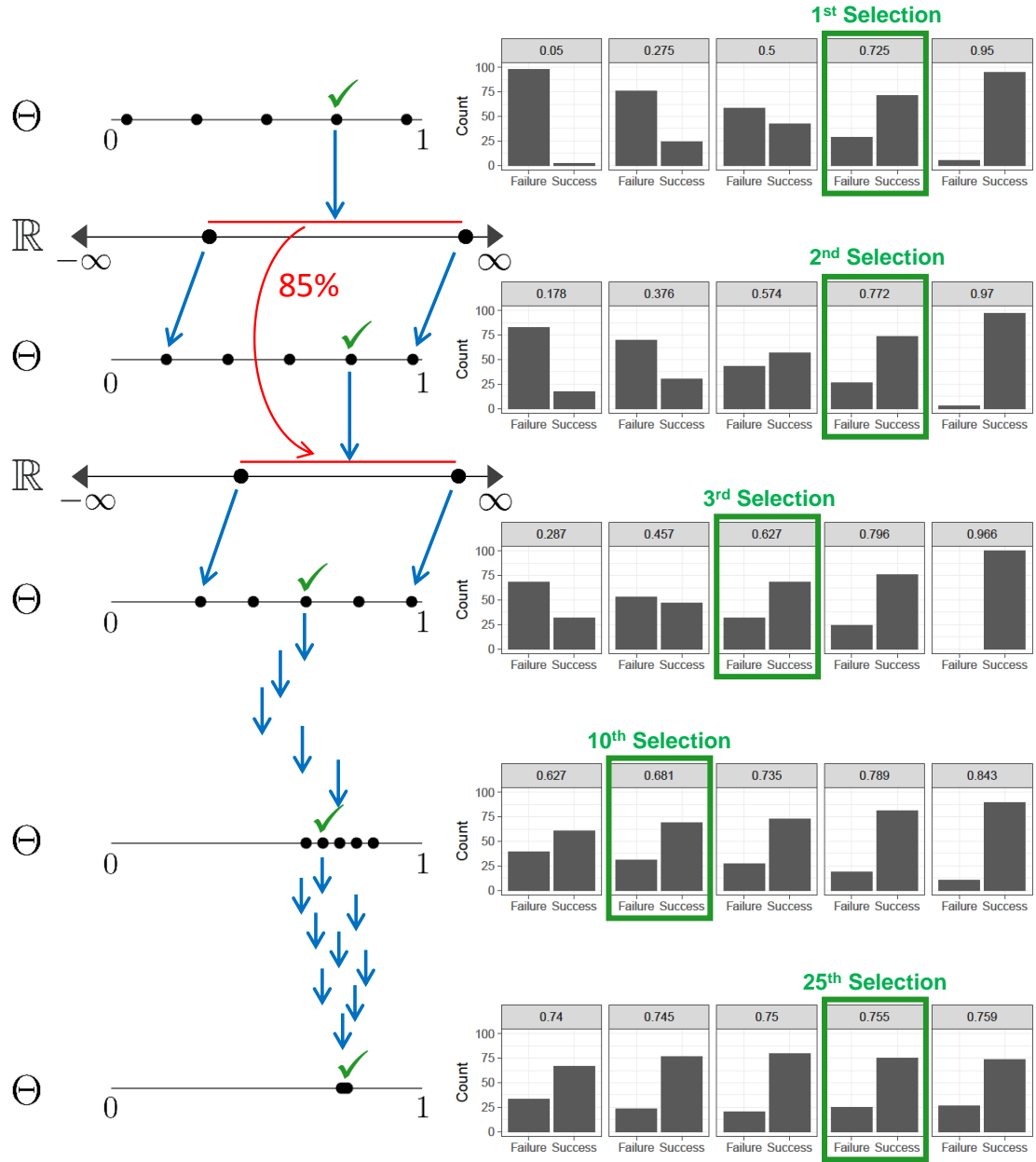


Figure 2.3. Algorithm steps for a Bernoulli data model. Check marks indicate the expert's selection.

tions with probabilities of success that correspond to the new proportions; bar charts of the data are again plotted; and the process continues, with the expert making 25 selections. For a Bernoulli data model, the tolerance threshold of 2% is reached after approximately 25 selections. Figure 2.3 displays a selection of graphs and their underlying proportions for a series of 25 selections.

Of the 25 total selections made, the final $K = 5$ are kept, with $\bar{\theta}_5^*$ representing their mean. Using $\bar{\theta}_5^*$ and the ESS of the beta family, the resulting system of equations for prior fitting is

$$\begin{aligned} \alpha + \beta &= \text{ess} = n \\ \frac{\alpha - 1}{\alpha + \beta - 2} &= \text{mode} = \bar{\theta}_5^*, \end{aligned}$$

which is solved with Equations (2.3) and (2.4) to provide a unique prior $\text{Beta}(\alpha, \beta)$.

2.4.2 $\mathcal{N}(\mu, \sigma^2)$ Data Model with μ and σ^2 Unknown

2.4.2.1 Description of the general scheme. We now present a graphical procedure for eliciting a joint prior when working with a $\mathcal{N}(\mu, \sigma^2)$ data model with σ^2 unknown. While various prior families exist for such a model, we elect to use the commonly-used Normal-inverse-gamma($\mu_0, \lambda, \alpha, \beta$) conjugate prior on $[\mu, \sigma^2]$. The hyperparameter space, \mathbf{H} , is now four-dimensional – twice the dimension of the hyperparameter space for the priors used previously with data models consisting of one unknown parameter. The two hyperparameters for those priors could be found directly from a combination of the expert’s selections and the sample size they specified. Now however, in addition to the sample size they specify, the expert makes selections corresponding to each parameter, only one of which varies for a given set of graphics, while the other is fixed at its previously-selected value. The procedure operates in a similar fashion to coordinate descent optimization. If the model is correct, the data generating mechanism is f_{θ^*} for some unknown $\theta^* = [\mu^*, (\sigma^2)^*]$. The goal is to

determine the mode of the joint prior – the most likely values the expert believes μ^* and $(\sigma^2)^*$ to be.

The first step of the scheme is to determine reasonable ranges $[l_\mu, u_\mu] \subset \mathbb{R}$ and $[l_{\sigma^2}, u_{\sigma^2}] \subset \mathbb{R}^+$ that contain μ^* and $(\sigma^2)^*$, respectively. Because both parameter spaces are unbounded, we cannot simply select values near the boundary, so we ask the expert to provide the smallest (largest) value an outcome could possibly be, denoted x_l (x_u), and select l_μ (u_μ) as the lowest (highest) μ for which x_l (x_u) is the 1st (99th) percentile. After applying the empirical rule, the initial variance, σ_0^2 , is set at $(x_u - x_l)/6$, and a reasonable initial range $[l_{\sigma^2}, u_{\sigma^2}]$ is found by dividing and multiplying σ_0^2 by 4.

With these initial parameter ranges in hand, we construct G -point equispaced meshes on $[l_\mu, u_\mu]$ and $[l_{\sigma^2}, u_{\sigma^2}]$. We then generate N pseudo-random variates from f_{μ, σ_0^2} , considering each of the G values of the mesh on $[l_\mu, u_\mu]$ as μ^* , and with σ^2 fixed at σ_0^2 . We construct histograms for those G scenarios and present them to the expert, who must select the one that would be most likely as a future dataset of N observations. Once a selection is made, the underlying mean used to generate the graphic, μ_1 , is recorded, and a slightly narrower range of parameter values is constructed around μ_1 . We next generate N pseudo-random variates from f_{μ_1, σ^2} , considering each of the G values of the mesh on $[l_{\sigma^2}, u_{\sigma^2}]$ as $(\sigma^2)^*$, and with μ fixed at μ_1 . Histograms are again constructed for the G scenarios, the expert chooses the best, and the σ^2 associated with the selected dataset is stored as σ_1^2 . G values in the range $[l_{\mu_1}, u_{\mu_1}]$ are taken for μ^* , with σ^2 fixed at σ_1^2 . Pseudo-random variates are once again generated and plotted, and the expert is asked for their opinion. This coordinate-descent-like process is repeated a number of times until each interval is a suitably small length, at which point the algorithm declares convergence. Figures 2.4 and 2.5 illustrate the first few steps of the procedure.

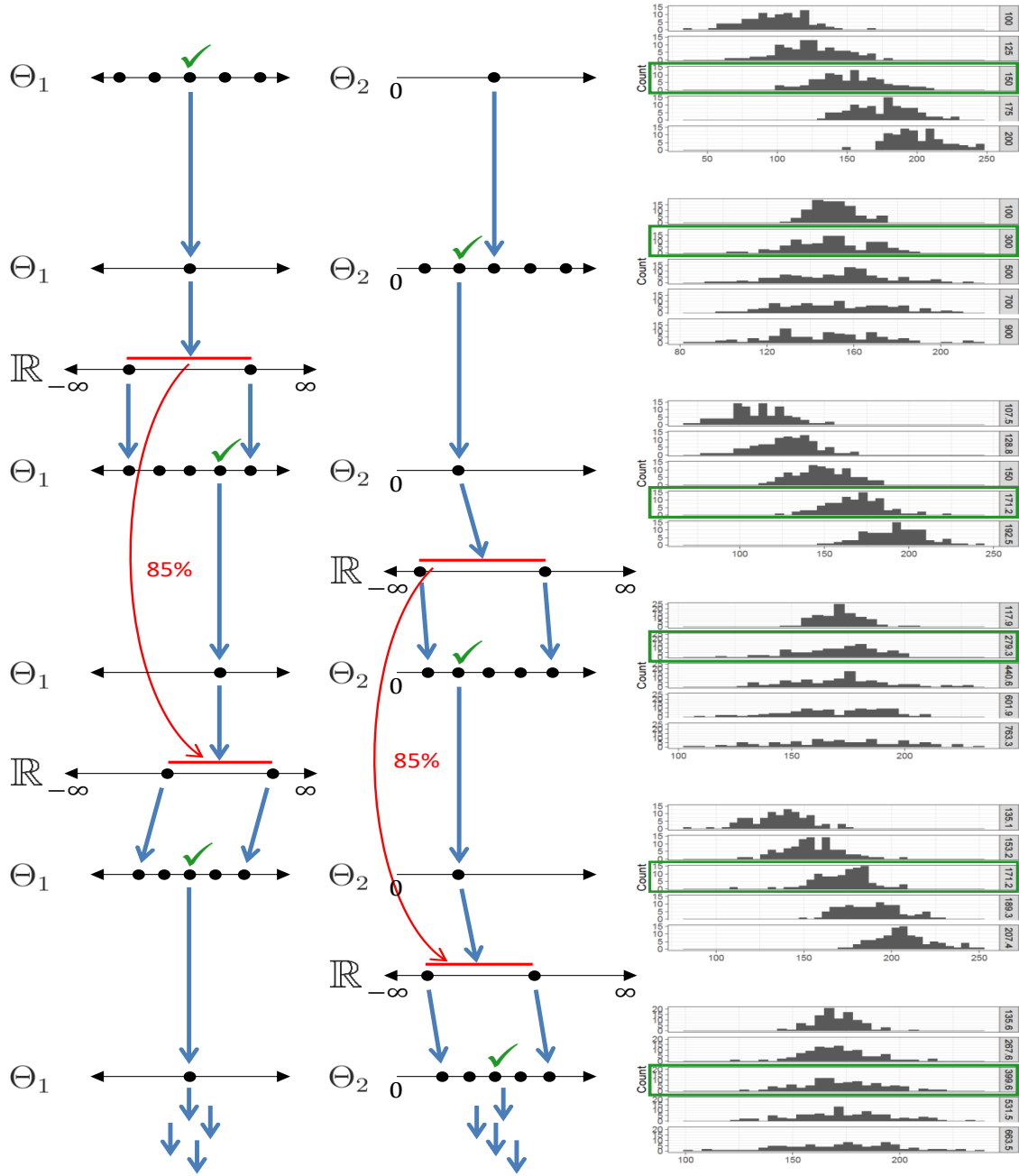


Figure 2.4. Algorithm steps for a $\mathcal{N}(\mu, \sigma^2)$ data model with μ and σ^2 unknown. Check marks indicate the expert's selection.

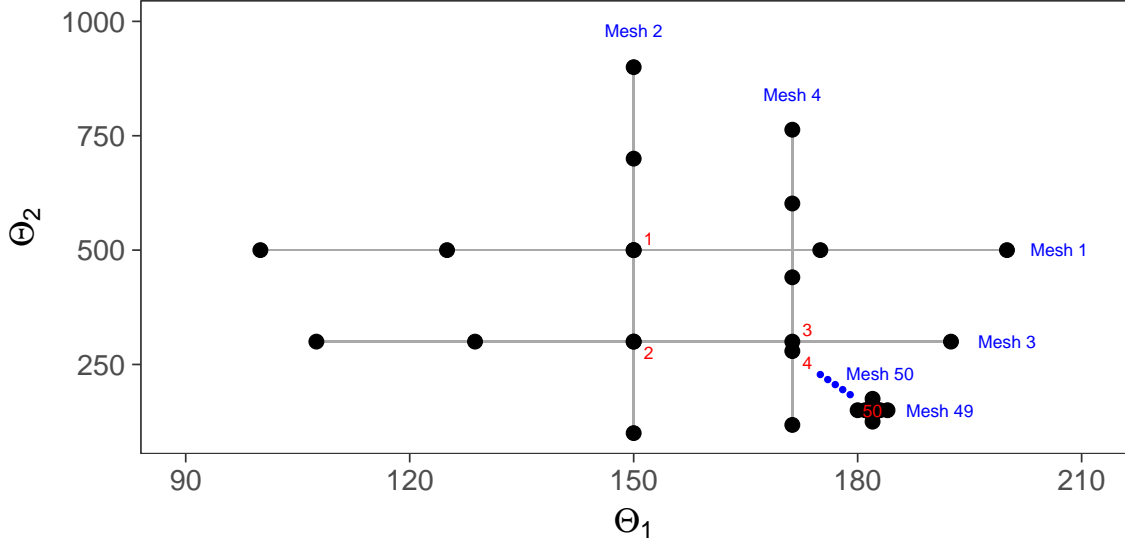


Figure 2.5. Two-dimensional view of the algorithm steps for a $\mathcal{N}(\mu, \sigma^2)$ data model with μ and σ^2 unknown. Numbers in red indicate the expert's selection.

The averages of the last K selections for each parameter, denoted $\bar{\mu}_K^*$ and $\bar{\sigma}_K^{2*}$, are taken to be the values the expert believes $[\mu^*, (\sigma^2)^*]$ most likely to be – the mode of $p_{\boldsymbol{\eta}}$, the joint prior on $[\mu, \sigma^2]$. We next consider the decomposition of $p_{\boldsymbol{\eta}}$ into the hierarchical prior structure

$$\begin{aligned} \mu | \sigma^2 &\sim \mathcal{N}(\mu_0, \sigma^2 / \lambda) \\ \sigma^2 &\sim \text{Inverse-Gamma}(\alpha, \beta). \end{aligned} \quad (2.5)$$

Setting $\text{ESS} = n$ for each of these priors, $p_{\boldsymbol{\eta}_1}$ and $p_{\boldsymbol{\eta}_2}$, as well as the modes equal to $\bar{\mu}_K^*$ and $\bar{\sigma}_K^{2*}$ then yields the nonlinear system

$$\text{ess}(p_{\boldsymbol{\eta}_1}) = n \quad (2.6)$$

$$\text{ess}(p_{\boldsymbol{\eta}_2}) = n \quad (2.7)$$

$$\text{mode}(p_{\boldsymbol{\eta}}) = \left[\frac{1}{K} \sum_{j=S-(K-1)}^S \mu_j^*, \frac{1}{K} \sum_{j=S-(K-1)}^S (\sigma_j^2)^* \right] \stackrel{\text{def}}{=} [\bar{\mu}_K^*, \bar{\sigma}_K^{2*}] \quad (2.8)$$

where $\{\mu_j^*\}_{j=1}^S$ and $\{(\sigma_j^2)^*\}_{j=1}^S$ are the S values for each parameter corresponding to the graphical selections made by the expert. A solution $\boldsymbol{\eta}^* = [\mu_0^*, \lambda^*, \alpha^*, \beta^*]$ to this

system is the vector of hyperparameters for the expert’s joint prior. The full process is detailed in algorithmic format in Algorithm 2.

2.4.2.2 Implementation notes. A few implementation notes of Algorithm 2 warrant further consideration.

First, after the expert makes a selection, the underlying parameter is transformed using an appropriate link function. Endpoints for the new interval are found on the link scale and then converted back to the parameter scale using the proper inverse-link function. Like before, this process ensures each parameter value falls within the correct support. Second, each new interval (on the link scale) is 85% of the length of the previous interval for each parameter. We set G and K each at 5 as was previously done.

While the overall framework of the method for the $\mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown is the same as for data models with one unknown parameter, an additional implementation choice had to be made for the former. Specifically, we had to address the question: how many parameters (one or both) should we allow to vary for each set of graphics to ensure reasonably accuracy for the resulting modes? Approaches based on various descent methods for optimization routines could be used to find these modes. However, using a method that allows both parameters to vary simultaneously at each step of the algorithm (e.g. a method that operates in a similar fashion to gradient descent) could make it difficult for the expert to choose the best graphic from the candidate set. For instance, suppose one of the G graphics appeared best in terms of its mean while another looked best in terms of its variability. While relatively quick to converge, a process based on a gradient descent scheme or any other that allows multiple parameters to vary simultaneously could likely lead to a prior that does not accurately reflect the expert’s opinion. We thus elected to use a coordinate descent-like scheme because it simplifies the selection process by allowing the expert

input : n – the number of previous observations
 N – the number of observations in a hypothetical future sample
 f_{θ} – the Normal data model parameterized by $\theta = [\mu, \sigma^2] \in \Theta$
 p_{η} – the Normal-inverse-gamma prior model parameterized by
 $\eta = [\mu_0, \lambda, \alpha, \beta] \in \mathbf{H}$
 r – the rate at which the window shrinks on the link scale (e.g. 15%)
 g – a link function $g : \Theta \rightarrow \mathbb{R}$ (typically the canonical link: e.g.
 $g(\mu) = \mu$ for the Normal μ ; $g(\sigma^2) = \log \sigma^2$ for the Normal σ^2)
 G – the mesh size/number of graphics to be presented (e.g. 5)
 K – the number of final selections to average for $\bar{\mu}_K^*$ and $\bar{\sigma}_K^{2*}$ (e.g. 5)
 tol – a tolerance for declaring convergence (e.g. interval width of 5
units on the parameter scale)

output: η^* , the hyperparameters for the prior on θ

```

1 for  $m \leftarrow 1$  to 2 do
2   | Construct interval  $[l_{\theta_m}, u_{\theta_m}] \subset \Theta_m$ 
3   |  $w_m \leftarrow g(u_{\theta_m}) - g(l_{\theta_m})$ 
4 end
5  $j \leftarrow 1$ 
6  $\theta_{2,0}^* \leftarrow (l_{\theta_2} + u_{\theta_2})/2$ 
7 repeat
8   | for  $m \leftarrow 1$  to 2 do
9     | for  $i \leftarrow 1$  to  $G$  do
10      |  $\theta_{m,i} \leftarrow l_{\theta_m} + (i - 1) \frac{u_{\theta_m} - l_{\theta_m}}{G - 1}$ 
11      | if  $m = 1$  then
12        | Sample  $\mathbf{d}^{(i)} = [d_1^{(i)} \dots d_N^{(i)}] \stackrel{iid}{\sim} f_{\theta_{m,i}, \theta_{m+1,j-1}^*}$ 
13      | else
14        | Sample  $\mathbf{d}^{(i)} = [d_1^{(i)} \dots d_N^{(i)}] \stackrel{iid}{\sim} f_{\theta_{m-1,j-1}^*, \theta_{m,i}}$ 
15      | end
16    | end
17    | Construct faceted (trellised) graphics with  $\mathbf{d}^{(1)}, \dots, \mathbf{d}^{(G)}$ 
18    |  $\theta_{m,j}^* \leftarrow$  the  $\theta_{m,i}$  corresponding to the plot of the expert's selection
19    |  $w_m \leftarrow (1 - r_m)w_m$ 
20    |  $(l_{\theta_m}, u_{\theta_m}) \leftarrow g^{-1}(g(\theta_{m,j}^*) \pm \frac{w_m}{2})$ 
21  | end
22  |  $j \leftarrow j + 1$ 
23 until  $u_{\theta_1} - l_{\theta_1} < tol_{\theta_1}$  and  $u_{\theta_2} - l_{\theta_2} < tol_{\theta_2}$ ;
24  $\eta^* \leftarrow$  Solution to Equations (2.6) – (2.8) for  $\eta$ 

```

Algorithm 2. General scheme of the procedure for a $\mathcal{N}(\mu, \sigma^2)$ data model with μ and σ^2 unknown.

to focus on only one parameter at a time when selecting the best graphic from each candidate set.

2.4.2.3 More on the prior inversion process. Eliciting a prior for a data model with two unknown parameters introduces additional levels of complexity not present when working with data models containing only one unknown parameter. For these latter models, the prior consisted of two hyperparameters, each of which could be found directly from the expert’s selections and the ESS they specified. For the $\mathcal{N}(\mu, \sigma^2)$ data model with σ^2 unknown, on the other hand, the expert makes selections that ultimately provide the mode of the joint prior. However, the expert-specified ESS and the mode translate to only three of the four hyperparameters, resulting in an underdetermined system of equations for prior fitting. How, then, does one obtain the remaining hyperparameter? The selections made by the expert provide no additional information than that already considered, but the ESS can provide more. In fact, this consideration leads to another important decision – which ESS does the sample size n correspond to: the ESS of the joint prior, the ESSs of the priors on $\mu|\sigma^2$ and σ^2 , or the combination of these? These quantities may differ, which leads to three cases for where to introduce the ESS warranting careful consideration.

The first case, which appears sound conceptually, proves difficult in practice. One approach is to find the set of hyperparameters that maximize the likelihood of the joint prior. To accomplish this, the problem becomes one of constrained optimization:

$$\begin{aligned} & \underset{\boldsymbol{\eta}}{\text{maximize}} && f(\boldsymbol{x}|\boldsymbol{\eta}) \\ & \text{subject to} && ess_{p_{\boldsymbol{\eta}}} = n, \end{aligned}$$

where $ess_{p_{\boldsymbol{\eta}}}$ represents the ESS of the joint prior, $p_{\boldsymbol{\eta}}$, and \boldsymbol{x} is the observation $(\bar{\mu}_K^*, \bar{\sigma}_K^{2*})$. While it is possible to implement a routine for such a problem, it suffers from an expensive computing time⁷ and relies on only one observation. Hence, case

⁷ In order to explore a reasonable number of points in the hyperparameter space, which is unbounded, a substantial number of non-trivial simulations must be run.

one is not a recommended approach for prior fitting.

The second case, which enforces the ESS on each distribution in the hierarchical prior structure specified in (5), is substantially quicker computationally, as it does not rely on a constrained optimization routine. Instead, it simply involves solving a system of equations. The first step in determining this system involves finding the mode of the joint Normal-inverse-gamma prior on $[\mu, \sigma^2]$. Doing so requires finding

$$\operatorname{argmax}_{\mu, \sigma^2} f(\mu, \sigma^2 | \mu_0, \lambda, \alpha, \beta),$$

where

$$f(\mu, \sigma^2 | \mu_0, \lambda, \alpha, \beta) = \frac{\sqrt{\lambda}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} e^{\frac{-2\beta + \lambda(\mu - \mu_0)^2}{2\sigma^2}}.$$

Taking the log of f yields

$$\begin{aligned} \log f &= -\frac{\log \sigma^2}{2} + \frac{1}{2} \log \left(\frac{\lambda}{2\pi} \right) + \log \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \right) - (\alpha + 1) \log \sigma^2 \\ &\quad - \frac{\beta}{\sigma^2} - \frac{\lambda(\mu - \mu_0)^2}{2\sigma^2}. \end{aligned}$$

Finding the partial derivatives of $\log f$ with respect to μ and σ^2 , and then setting each partial equal to 0 leads to the nonlinear system

$$\begin{aligned} \frac{\partial \log f}{\partial \mu} &= -\frac{\lambda(\mu - \mu_0)}{\sigma^2} \stackrel{set}{=} 0 \\ \frac{\partial \log f}{\partial \sigma^2} &= \frac{2\beta + \lambda(\mu - \mu_0)^2}{2\alpha + 3} \stackrel{set}{=} 0. \end{aligned}$$

The solution to this system, which is the mode of the joint Normal-inverse-gamma prior, is

$$\operatorname{argmax}_{\mu, \sigma^2} f(\mu, \sigma^2 | \mu_0, \lambda, \alpha, \beta) = \left[\mu_0, \frac{2\beta}{2\alpha + 3} \right]. \quad (2.9)$$

We must now verify the solution found in (2.9) is, in fact, a maximum. We first find

the second-order partial derivatives of $\log f$ with respect to μ and σ^2 to be

$$\frac{\partial^2 \log f}{\partial \mu^2} = -\frac{\lambda}{\sigma^2} \quad (2.10)$$

$$\frac{\partial^2 \log f}{\partial (\sigma^2)^2} = \frac{\sigma^2(2\alpha + 3) - 4\beta - 2\lambda(\mu - \mu_0)^2}{2\sigma^6} \quad (2.11)$$

$$\frac{\partial^2 \log f}{\partial \mu \partial \sigma^2} = \frac{\lambda(\mu - \mu_0)}{\sigma^4} \quad (2.12)$$

$$\frac{\partial^2 \log f}{\partial \sigma^2 \partial \mu} = \frac{2\lambda(\mu - \mu_0)}{2\alpha + 3}. \quad (2.13)$$

The Hessian matrix is thus

$$H(\mu, \sigma^2) = \begin{bmatrix} -\frac{\lambda}{\sigma^2} & \frac{2\lambda(\mu - \mu_0)}{2\alpha + 3} \\ \frac{\lambda(\mu - \mu_0)}{\sigma^4} & \frac{\sigma^2(2\alpha + 3) - 4\beta - 2\lambda(\mu - \mu_0)^2}{2\sigma^6} \end{bmatrix}. \quad (2.14)$$

Evaluating (2.10) at the solution found in (2.9) leads to

$$\left. \frac{\partial^2 \log f}{\partial \mu^2} \right|_{\mu=\mu_0, \sigma^2=\frac{2\beta}{2\alpha+3}} = -\frac{\lambda(2\alpha + 3)}{2\beta} < 0. \quad (2.15)$$

The second-order partial derivative in (2.15) is always negative, as λ , α and β are all positive. Next, evaluating the Hessian in (2.14) at the solution in (2.9) yields

$$H(\mu, \sigma^2) \big|_{\mu=\mu_0, \sigma^2=\frac{2\beta}{2\alpha+3}} = \begin{bmatrix} -\frac{\lambda(2\alpha+3)}{2\beta} & 0 \\ 0 & -\frac{(2\alpha+3)^3}{8\beta^2} \end{bmatrix}.$$

The Jacobian of the second-order partial derivatives is thus

$$\begin{aligned} J &= \left| H(\mu, \sigma^2) \big|_{\mu=\mu_0, \sigma^2=\frac{2\beta}{2\alpha+3}} \right| \\ &= \begin{vmatrix} -\frac{\lambda(2\alpha+3)}{2\beta} & 0 \\ 0 & -\frac{(2\alpha+3)^3}{8\beta^2} \end{vmatrix} \\ &= \frac{\lambda(2\alpha + 3)^4}{16\beta^3} > 0. \end{aligned} \quad (2.16)$$

The Jacobian in (2.16) is always positive, as λ , α and β are all positive. Finally, because at least one second-order partial derivative evaluated at (2.9) is negative,

and the Jacobian of the Hessian evaluated at (2.9) is positive, $\left[\mu_0, \frac{2\beta}{2\alpha+3}\right]$ is, in fact, a maximum.

Now recall the expert's selections provide the mode of the joint prior:

$$\operatorname{argmax}_{\mu, \sigma^2} f(\mu, \sigma^2 | \mu_0, \lambda, \alpha, \beta) = \left[\bar{\mu}_5^*, \bar{\sigma}_5^{2*}\right]. \quad (2.17)$$

Setting (2.9) equal to (2.17) then leads to $\mu_0 = \bar{\mu}_5^*$ and $\beta = \bar{\sigma}_5^{2*}(\alpha + \frac{3}{2})$. Next, the ESS for the $\mathcal{N}\left(\mu_0, \frac{\sigma^2}{\lambda}\right)$ prior on $\mu | \sigma^2$ is λ (Jackman, 2009). Similarly, after reparameterizing the prior on σ^2 as inverse-gamma $\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$, this prior's ESS is ν_0 . The resulting system of equations for prior fitting is thus

$$\mu_0 = \bar{\mu}_5^* \quad (2.18)$$

$$\lambda = \text{ess} \quad (2.19)$$

$$\alpha = \text{ess}/2 \quad (2.20)$$

$$\beta = \bar{\sigma}_5^{2*}(\text{ess} + 3)/2. \quad (2.21)$$

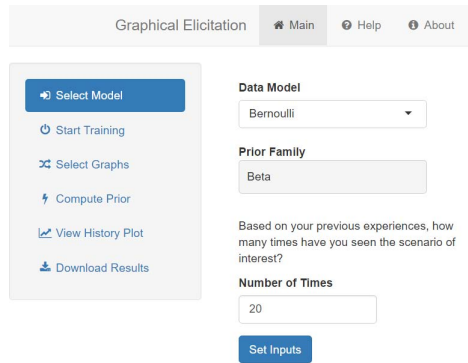
For the third case, which is a combination of cases one and two, the ESS is enforced on the joint Normal-inverse-gamma $(\mu_0, \lambda, \alpha, \beta)$ prior in addition to its hierarchical parts: the $\mathcal{N}\left(\mu_0, \frac{\sigma^2}{\lambda}\right)$ prior on $\mu | \sigma^2$ and the inverse-gamma (α, β) prior on σ^2 . Optimization remains a difficult task computationally, and the existence of a solution is not guaranteed.

We thus proceeded with the approach from case two. However, while this approach successfully assists in the elicitation of a prior, caution must still be exercised. After finding the hyperparameters that satisfy Equations (2.18) through (2.21), the joint prior's ESS should then be computed and compared to the sample size provided by the expert in order to ensure the elicited prior is not substantially more informative than intended.

2.5 Implementation

Shiny app implementations for all four data models described in Section 2.4 are freely available from `ccasement.shinyapps.io/graphicalElicitation` (Chang et al., 2016). The source code for these apps can be found on GitHub at `github.com/ccasement/graphicalElicitation`. In this section we demonstrate a concrete use of the graphical prior elicitation method through a series of screenshots of the app. To demonstrate the method’s ability to accurately quantify expert opinion, we use the application to elicit a beta prior for a Bernoulli proportion.

To initialize the process, suppose the expert specifies they have seen the scenario of interest twenty times before, as can be seen in Figure 2.6.



The screenshot shows the 'Graphical Elicitation' Shiny app interface. At the top, there is a navigation bar with 'Main', 'Help', and 'About' tabs. The 'Main' tab is active. On the left side, there is a sidebar with a list of actions: 'Select Model' (highlighted with a blue button), 'Start Training', 'Select Graphs', 'Compute Prior', 'View History Plot', and 'Download Results'. The main content area on the right contains the following inputs: 'Data Model' is a dropdown menu set to 'Bernoulli'; 'Prior Family' is a text input field set to 'Beta'; a text prompt reads 'Based on your previous experiences, how many times have you seen the scenario of interest?'; 'Number of Times' is a text input field set to '20'; and a blue 'Set Inputs' button is at the bottom.

Figure 2.6. Expert inputs to the Shiny app for graphical elicitation of the Bernoulli p .

The Shiny application then trains the expert to understand the natural variability inherent in real data sets using the Rorschach procedure described in Section 2.3.1. Nine bar charts are plotted next to one another, each for a different random sample of size 50 from a Bernoulli($p = 0.5$) distribution, where the $p = 0.5$ is based on the expert-input 50 expected successes out of a sample of size 100. This can be seen in Figure 2.7. These graphs allow the expert to visualize expected and unexpected behavior for data from such a model. (The Bernoulli case is not very surprising.)

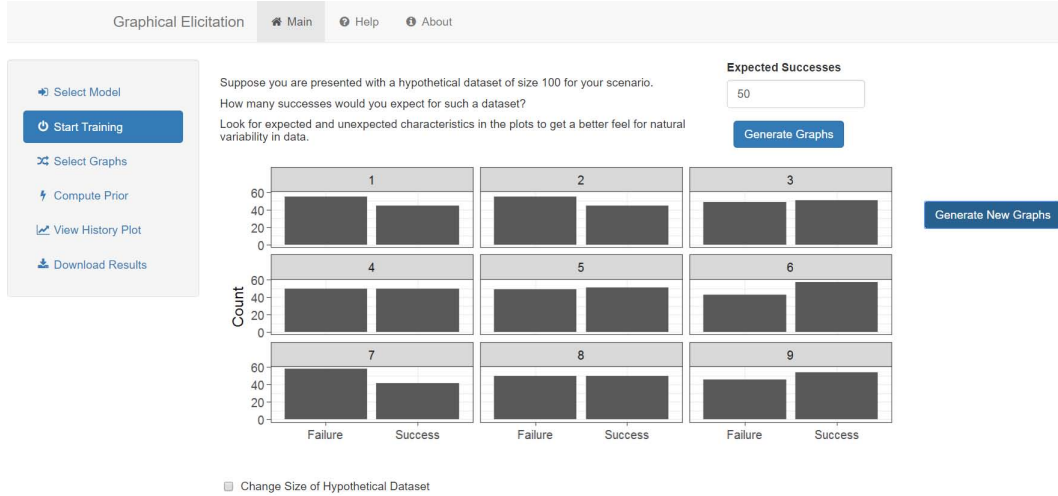


Figure 2.7. Rorschach training plots.

Once the expert feels comfortable visualizing natural variability, he/she begins the selection process. Five bar charts, based on the initial mesh of $\{0.05, 0.275, 0.5, 0.725, 0.95\}$ discussed in Section 2.4, are presented in a randomized order, as can be seen in Figure 2.8.⁸ The expert is then tasked with selecting the most likely graphic from the candidates, after which a new set of five bar charts is generated and plotted.

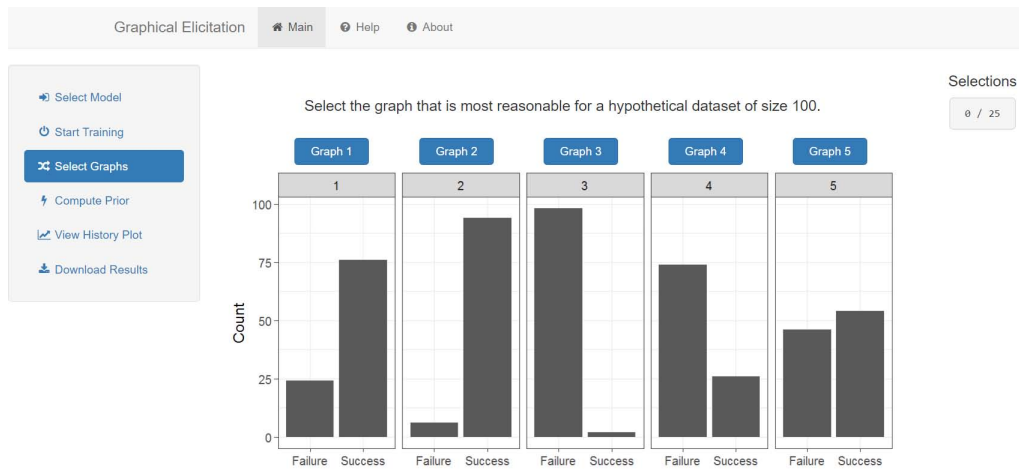


Figure 2.8. Example of graphs presented to the expert at each step.

⁸ Histograms are used for other data models.

After the expert has made 25 selections, the prior is computed. Figure 2.9 displays information about the prior that is provided to the expert, such as: (1) the prior family, (2) the elicited parameters, (3) summaries of the elicited prior, and (4) a density plot of the elicited prior. These plots and summaries enable the expert to further assess the adequacy of the prior distribution they helped elicit interactively, and pdf's can be exported for download. Additionally, the expert may adjust the sample size they provided at the onset of the process so that they may examine its impact on the variability of the prior.

The application also displays a history plot (Figure 2.10), which shows the proportions at each step: the selected proportions (red) and the four unselected proportions (black).

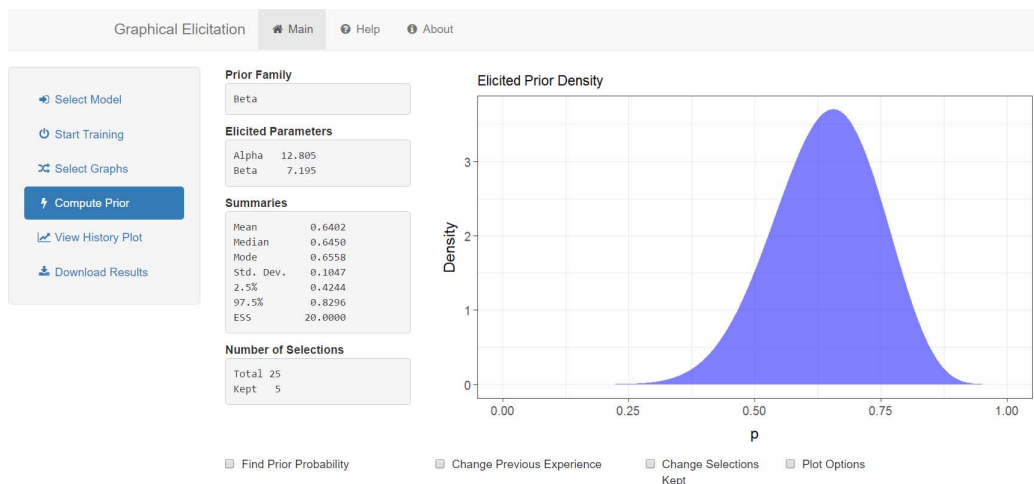


Figure 2.9. Elicited prior and its summaries.

2.6 Conclusion

While current elicitation methods enable experts to incorporate their beliefs regarding a parameter into a prior distribution, they suffer from a number of drawbacks. In this paper we have proposed interactive graphical methods that provide an alternative to the standard analytical methods by refining the elicitation process in each of the areas of their shortcomings. First, the methods improve the accuracy of

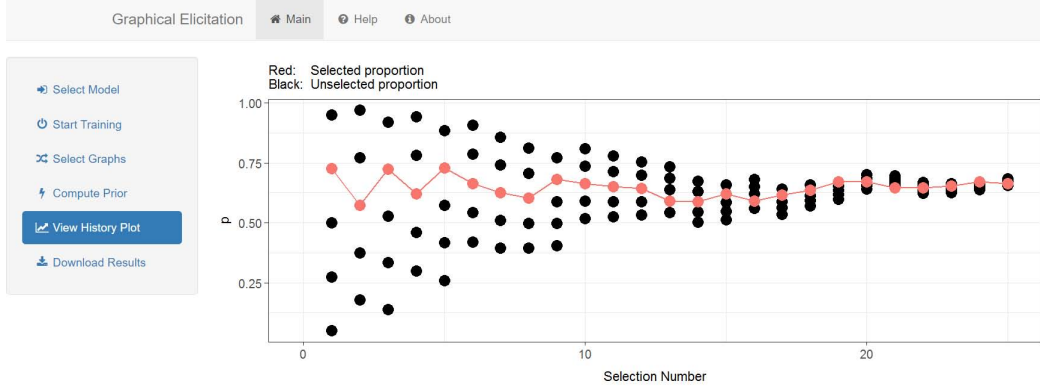


Figure 2.10. History plot of the proportions at each step.

the elicited prior by allowing the expert to work on the most observational of scales – the data itself – rather than relying on the expert’s ability to accurately quantify summaries for the parameter of interest. The method also addresses each stage of the elicitation process in a holistic way, recognizing and emphasizing the process over its constituent parts through the use of the freely-accessible, online Shiny app. The process is also extensible to data models not discussed in this chapter.

Yet, while the proposed graphical scheme presents clear strengths, it relies on the equivalent prior sample (EPS) method, a related analytical elicitation strategy which studies have shown does not perform as well as other elicitation methods due to its tendency to underestimate the variability of the elicited prior (Schaefer and Borchering, 1973, Winkler, 1967). To address this issue, we allow the expert to examine the effect of changing the ESS on the variability of the prior through the use of a slider in the Shiny app. Even still, it is not clear the graphical strategy will incur the same challenges.

CHAPTER THREE

Stochastic Procedures for Graphical Prior Elicitation in Univariate Models

Abstract

Current methods for prior elicitation call for expert belief in the form of numerical summaries. However, certain challenges remain with such strategies. Drawing on the Metropolis algorithm in addition to recent advances made in graphical inference, we propose an interactive scheme for prior elicitation in which experts work directly with graphics instead of parameters. An expert is presented two hypothetical future datasets in the form of graphics and makes a selection regarding their relative likelihood. The full process, which is stochastic in its underlying structure, mimics Metropolis. Using the general scheme, we develop procedures for data models used regularly in practice: Bernoulli, Poisson, and Normal, though it extends to additional univariate data models as well. A free, open-source Shiny application designed for these procedures is also available online.

3.1 Introduction

Eliciting a prior for a Bayesian analysis demands particular care be taken due to a prior's ability to strongly impact the analysis and potentially undermine the accuracy of the results. To this end, substantial literature has been written on prior elicitation from members of wide-ranging academic communities. Garthwaite et al. (2005) provides an overarching summary of the literature while also detailing a four-stage process for prior elicitation that consists of the following stages: setup, elicitation, fitting, and adequacy assessment.

During the setup stage, the prior family, the elicitation method to be used, and the necessary statistical summaries to be acquired from the expert(s) are all

determined, and the expert(s) is selected and trained. During the elicitation stage, the expert quantifies their knowledge in the form of the desired summaries (e.g. mean and percentiles). During the fitting stage, the statistician converts the summaries into the hyperparameters of a distribution from the prior family chosen previously, and the expert is then asked to assess whether the resulting prior accurately reflects their belief during the adequacy assessment stage.

While many methods and tools have been developed that assist with the elicitation process, certain drawbacks remain. One issue is that it can be difficult for experts to provide reliable estimates, while another is the undervaluing of the elicitation process as a whole, myopically focusing on the second and third stages.

In this chapter we turn to advances made in the areas of statistical graphics and information visualization to propose a novel graphical procedure for prior elicitation that addresses these drawbacks. The resulting procedure promotes good practice by enforcing the elicitation process as a synergistic whole. The process starts with the proper training of the expert using the Rorschach procedure (see Section 3.3). Then, rather than having to specify summaries directly, the expert makes a series of selections of graphics they believe to be potential hypothetical future datasets, with the parameters underlying the selected graphics converted to the hyperparameters of the chosen prior family. The elicited prior and its summaries are then immediately presented to the expert for verification. Free web-based Shiny implementations for four commonly-used models – Bernoulli with a beta prior, Poisson with a gamma prior, Normal (σ^2 known) with a Normal prior, and Normal (σ^2 unknown) with a Normal-inverse-gamma prior – are available at `ccasement.shinyapps.io/graphicalElicitationMCMC`.

The chapter proceeds as follows. In Section 3.2 we discuss the prior elicitation process, such as existing methods, their advantages and disadvantages, and tools for working with them. In Section 3.3 we present two graphical procedures developed for

statistical inference – the Rorschach and line-up – which we draw on for the proposed elicitation method. The proposed method, a variation on a common stochastic theme, follows in Section 3.4. We then provide a demonstration with screenshots of the Shiny application in Section 3.5. In Section 3.6 we discuss additional theoretical elements of the proposed method before summarizing the chapter in Section 3.7.

3.2 *Prior Elicitation*

3.2.1 *Existing Methods and Implementations*

Many prior elicitation methods have been developed, nearly all of which are analytical in the following sense. Experts are asked to provide summaries of a parameter(s), which are then converted into the hyperparameters of a predetermined prior family. Examples of such methods, which include the mode and percentile method, the probability density function method, the cumulative distribution function method, and the equivalent prior sample method, among others, are detailed in Garthwaite et al. (2005), Kahle et al. (2014), and O’Hagan et al. (2006).

To assist with the elicitation process, some of the methods have been implemented in software packages and other tools. When using these tools, the expert is asked to specify summaries appropriate for a given method, and the conversions, which can be relatively complicated, are performed behind the scenes. Examples of such tools include:

- (1) SHELF (the Sheffield Elicitation Framework), which runs in R (Oakley and O’Hagan, 2010),
- (2) the MATCH (the Multidisciplinary Assessment of Technology Centre for Healthcare) Uncertainty Elicitation Tool, which is based on SHELF but runs through a web browser rather than directly through R (Morris et al., 2014),
- (3) BetaBuster, which operates through a Java applet (Su, 2006), and

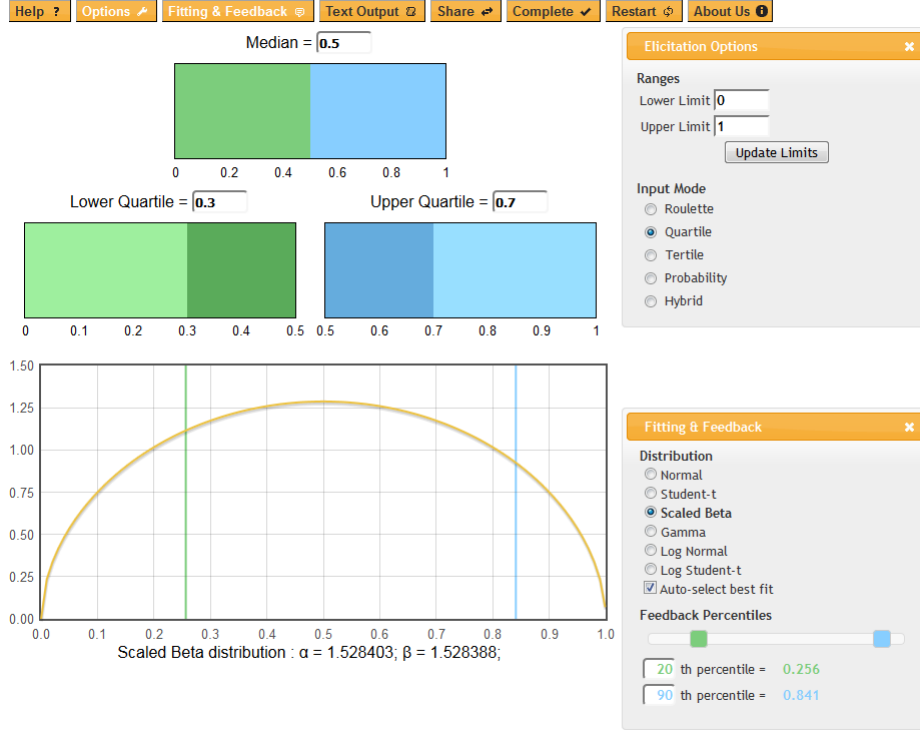


Figure 3.1. The MATCH Uncertainty Elicitation Tool is a free browser-based JavaScript elicitation tool.

(4) the Wolfram beta elicitation tool (Kahle et al., 2014).

Another tool, which is graphical and more exploratory in nature than the four discussed previously, operates in Microsoft Excel (Jones and Johnson, 2014). Screenshots of two of these tools – MATCH and BetaBuster – are shown in Figures 3.1 and 3.2.

3.2.2 Shortcomings of Existing Methods

While the methods discussed in the previous section are important contributions that have enabled the elicitation of priors, they also contain certain shortcomings. In this section we describe two such concerns.

One concern is the expert’s ability to accurately quantify summaries that represent their belief in a parameter. The elicitation method employed may be sensitive to changes in the summaries, which could, in turn, undermine the accuracy of an

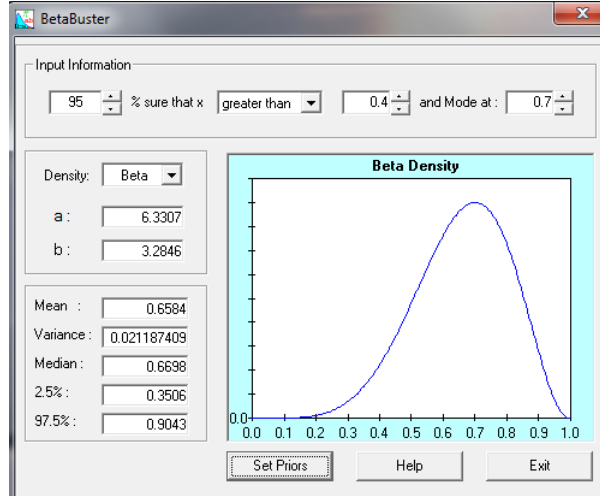


Figure 3.2. BetaBuster is a free **Java** application used for mode-percentile elicitation of beta distributions.

analysis, as prior distributions play an important role in determining posteriors. For instance, when using the mode and percentile method, small changes in the value of a percentile can dramatically alter how informative the resulting prior is (Blair, 2017).

Undervaluing the elicitation process as a whole is a second concern. The statistician must always ensure the expert is properly trained prior to the elicitation stage and that they carefully perform a post-elicitation assessment of the prior after the fitting stage. Many studies have concluded that people often exert overconfidence in their beliefs that directly results in estimates that are overly precise (see, for example, Keren, 1991; Lichtenstein et al., 1977; Lichtenstein and Fischhoff, 1977; and Oskamp, 1965). Such estimates jeopardize the accuracy of analyses if unduly informative priors are used, especially if the corresponding belief is incorrect (in the sense that it is not concentrated on the true unknown parameter value).

3.3 Graphical Procedures for Statistical Inference

We now briefly turn our attention to two graphical procedures for inference: the Rorschach and line-up procedures (Buja et al., 2009; Wickham et al., 2010). After a

discussion of both, we draw on them for the graphical elicitation methods proposed in Section 3.4.

3.3.1 *Rorschach Procedure*

The Rorschach procedure is used to train an individual to better understand the natural variability that exists in data models. When performing the procedure, G datasets of reasonable size are randomly generated from the same distribution, called the “null” distribution (Wickham et al., 2010), and graphs of the datasets are plotted next to one another. The individual examines and compares the graphs, focusing on features they do and do not expect given the particular null distribution. Those features, both expected and unexpected, can be attributed to the natural variability of the underlying null distribution for the sample size selected. In fact, for a fixed sample size, the data model induces a distribution on the set of graphics, and the Rorschach procedure allows an expert to learn this distribution through a series of observations. So long as the graphing method is sufficiently granular with respect to the sample space of the data, the two can be considered equivalent.

For instance, suppose an individual wants to become more comfortable visualizing the natural variability in samples of size 500 from the Beta(0.5, 0.5) distribution. The Rorschach procedure would entail randomly generating G datasets, each of size $n = 500$ from the Beta(0.5, 0.5) distribution, and then plotting a histogram of each side-by-side. The process can be repeated many times to more fully learn the natural variability. Figure 3.3 demonstrates such an example for $G = 16$.

Natural variability clearly appears in each graph, as none of the sixteen histograms shows a perfectly round bathtub shape. The Rorschach procedure thus improves an individual’s ability to understand such variability for the Beta(0.5, 0.5) distribution and decreases their chance of concluding a dataset does not come from that distribution simply because its graph does not appear exactly as expected. While

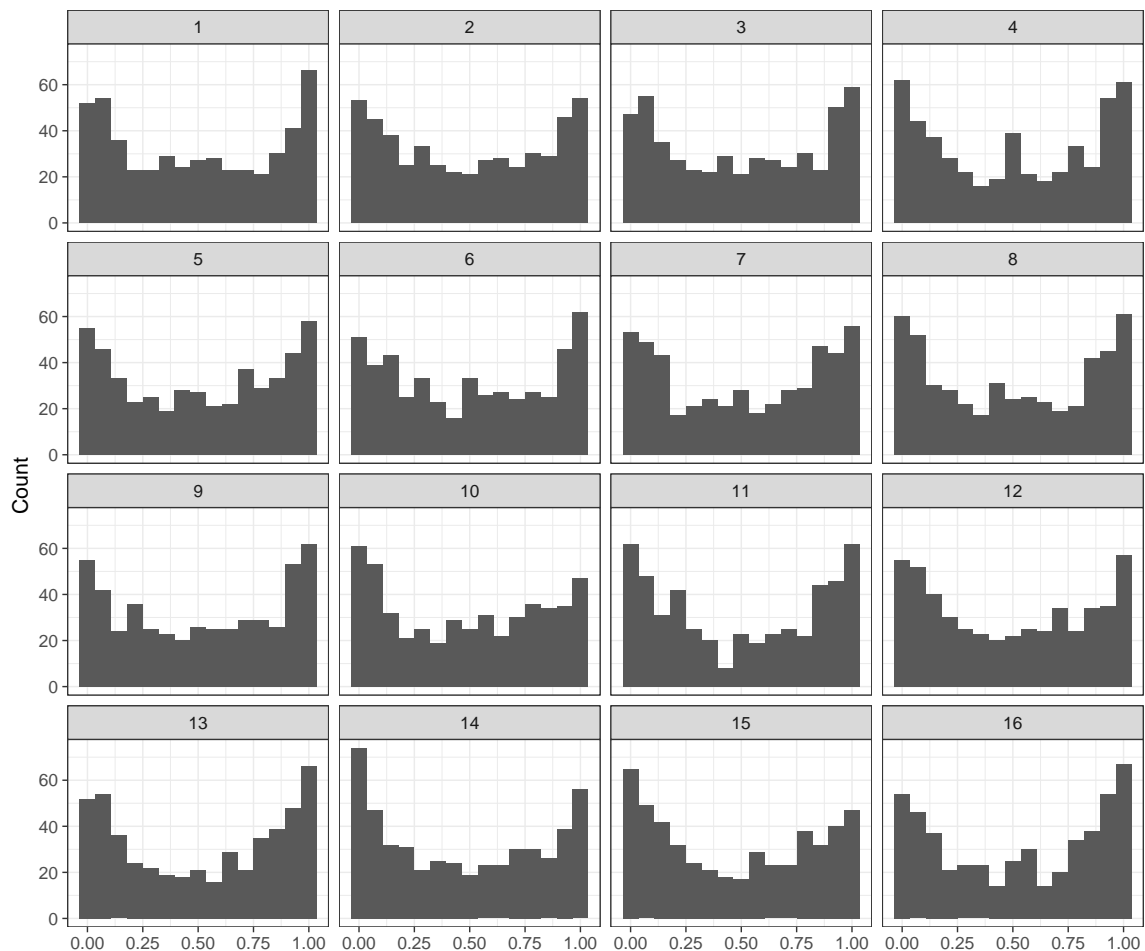


Figure 3.3. Sixteen histograms, each of 500 random observations from the $\text{Beta}(0.5, 0.5)$ distribution.

the Rorschach procedure involves the inspection of graphics without any selections made (e.g. best or worst), the next procedure described, the line-up, does require a selection be made. Thus, the Rorschach is a natural precursor to the line-up.

3.3.2 Line-up Procedure

The line-up procedure, a second graphical inference procedure, acts as a graphical goodness-of-fit test. The null hypothesis specifies that the data come from a particular distribution, and the alternative hypothesis specifies that the data do not come from that distribution. If testing a model using the line-up procedure, the

model is fit to the data and the resulting distribution treated as the null. Next, $G - 1$ datasets, all of the same size as the initial dataset, are randomly generated from the null distribution, and graphs are plotted next to one another, with the plot of the real dataset randomly mixed in. The individual is then asked to select the graph of the real data from the G graphs. If they successfully choose the real dataset, the null hypothesis is rejected, suggesting the real data do not come from the null distribution.

For example, suppose an individual wants to test the homogeneity of variance assumption in a linear regression model. While analytical tests have been developed for such a test, the line-up procedure could be utilized as well by visually examining residual plots. In fact, an initial graphical inspection of the residuals is a standard step when testing for homogeneity of variance of the errors. To use the line-up procedure for such a test, a linear model is first fit to the data, and the residuals are calculated. $G - 1$ datasets, all of the same size as the initial dataset, are randomly generated from the fitted model, and a residual plot is created for each. The residual plot for the real dataset is then randomly mixed in with the residual plots for the $G - 1$ simulated datasets. The individual must next select the plot that corresponds to the real data. If they successfully do so, then the null hypothesis of homogeneity of variance of the errors is rejected. Figure 3.4 illustrates an example for $G = 16$ datasets, each of size $n = 50$. For the solution to the example, see footnote ¹.

The two graphical inference procedures described here – the Rorschach and line-up – are well suited for the graphical elicitation methods described in the following section for two main reasons. First, the procedures are accessible to non-statisticians, as they do not require a strong statistical background; generally experts are familiar with common statistical graphics. Second, the line-up – especially after the individual is trained using the Rorschach – allows them to visually determine whether a difference between graphics is meaningful based on their expertise.

¹ Graph 12 displays the real data.

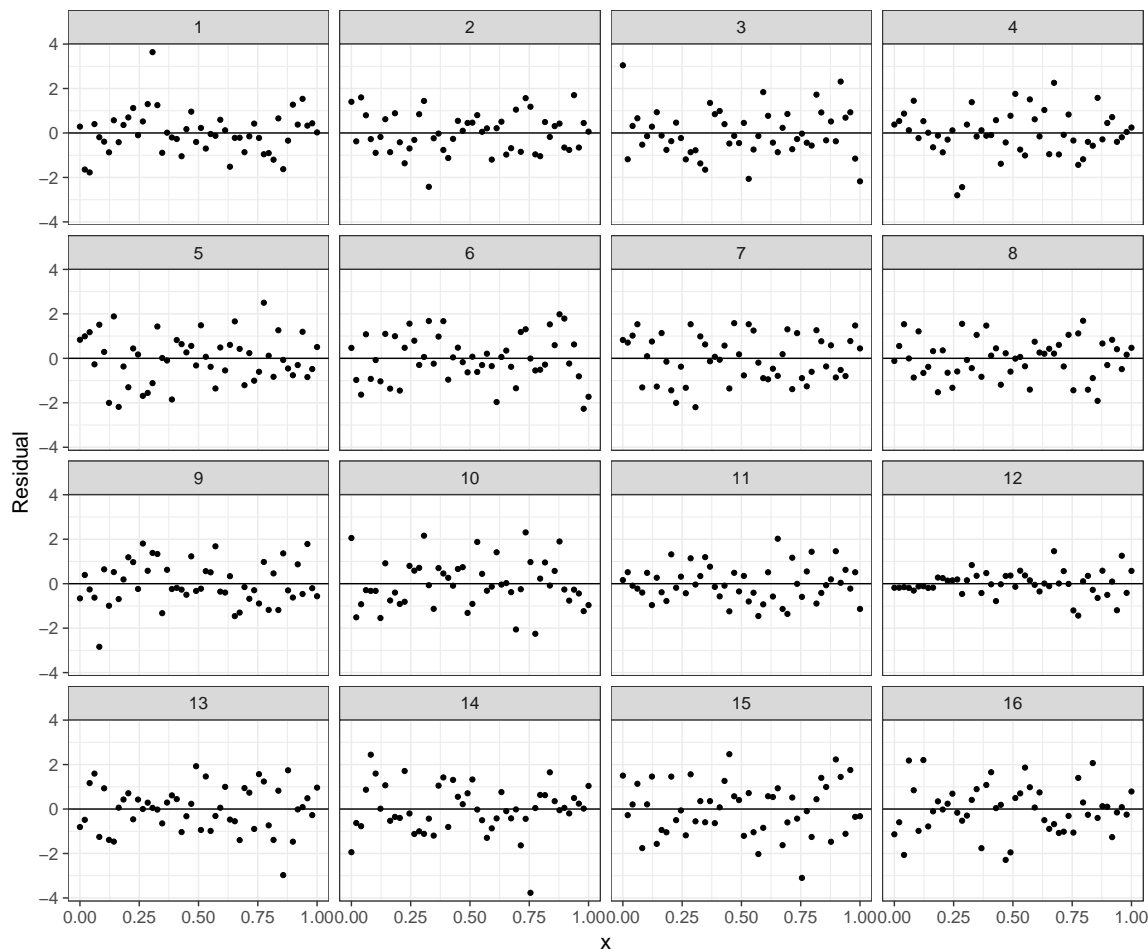


Figure 3.4. Fifteen scatterplots of simulated residuals, one of real residuals randomly mixed in.

3.4 Graphical Elicitation

We now present a graphical procedure for prior elicitation built on top of a stochastic infrastructure with a Metropolis foundation for transitioning from one set of graphics to the next. Like the deterministic method in Chapter 2, this procedure addresses the shortcomings of the analytical methods discussed in Section 3.2. For the proposed method, the expert makes a series of selections of graphics that are displayed in a lineup-style fashion, after which a prior is fit to the accepted parameters underlying the graphics. We first describe a method for univariate data models

with one unknown parameter and then present a similar method for the $\mathcal{N}(\mu, \sigma^2)$ data model with both μ and σ^2 unknown.

3.4.1 Data Models with One Unknown Parameter

3.4.1.1 Description of the procedure. Suppose a data model f_θ and a prior p_η on θ , where $\eta = [\eta_1, \eta_2]$, are assumed.² In the first step of the procedure, the expert is asked to provide a typical measurement value for a hypothetical future dataset of size N . This value x determines the initial step $\theta^{(0)}$ of the algorithm and helps calibrate it. For example, for a Bernoulli data model the initial parameter $p^{(0)}$ is set at x/N , where x represents the expert-specified typical number of successes for a hypothetical future dataset. For a Poisson model or a Normal data model with σ^2 known, the initial parameter is set at $\lambda^{(0)} = x$ and $\mu^{(0)} = x$, respectively. The expert then goes through Rorschach training, learning expected and unexpected characteristics of samples of size N from Bernoulli distributions. The expert is able to specify the probability and even regenerate graphics for further exploration.

After the expert is comfortable with their training, they move on to the graphical generation and selection process. A proposed parameter value θ^{prop} is drawn from a $\mathcal{N}(\theta^{(0)}, \sigma_\theta)$ proposal distribution (a Normal proposal is used for all the data models discussed). Next, N probabilities $u_i \in \mathbf{u}$ are randomly drawn from a $\text{Unif}(0, 1)$ distribution and are converted to observations from the distribution of interest using inverse transform sampling. This results in two datasets:

$$\begin{aligned} \mathbf{x}^{\text{current}} &= F_{\theta^{(0)}}^{-1}(\mathbf{u}), \text{ and} \\ \mathbf{x}^{\text{proposed}} &= F_{\theta^{\text{prop}}}^{-1}(\mathbf{u}). \end{aligned}$$

This conversion allows for a direct comparison of the current and proposed parameter values based on a single original sample. Graphically, this amounts to the same

² The procedure extends to hyperparameter spaces of dimension greater than two, but we focus on two-dimensional spaces.

dataset being either shifted or scaled (with possible re-binning for histograms) depending on the parameter of interest. Such a choice is used to best maintain the Metropolis foundation seen shortly.

The current and proposed datasets $\mathbf{x}^{\text{current}}$ and $\mathbf{x}^{\text{proposed}}$ are then presented to the expert in the form of two graphics, and the expert is tasked with making a selection from a set of five options: the proposed plot is more likely than the current, both plots are equally likely, or three options representing how much more likely the current plot is than the proposed plot. If the expert selects the option where the proposed graphic is more likely, then the proposed step is accepted as the new current step of the sampler. If the expert specifies that both graphics are equally likely, again the proposed step is immediately accepted. These automatic acceptances mimic that of a Metropolis sampler – when the proposed step is equally likely or more likely than the current step under the target distribution, the proposed is accepted with probability one. On the other hand, if the expert feels the current graphic is more likely than the proposed, then they must choose how much more likely the graphic is. In the Metropolis algorithm, this value is used to determine the transition probability – the probability of accepting the proposed value as having come from the target distribution. Since the specification of such a precise number by an expert seems both implausible and cumbersome, we restrict the odds to be one of three values: $\mathbf{O} = \{3, 25, 10^6\}$. The expert’s choice of odds then determines the probability with which the proposed value will be accepted, as the odds and acceptance probability α are in one-to-one correspondence through the relationship

$$\alpha = \frac{1}{O_i^{\text{selected}}}.$$

The three odds options, as well as their consequences, are discussed in detail in Chapter 4. The corresponding Metropolis-like algorithm is referred to as rigid Metropolis or rigid MCMC (Markov chain Monte Carlo) for short.

The expert is asked to make a total of M selections, with M large enough such that the algorithm has had sufficient time to explore the parameter space. We have found $M = 100$ to work well in practical scenarios and simulations. The MLEs³ $\hat{\eta}_{1_{\text{MLE}}}$ and $\hat{\eta}_{2_{\text{MLE}}}$ for the hyperparameters of the prior of interest are then computed based on the chain of accepted parameter values underlying the graphics. Algorithm 3 lays out the procedure algorithmically.

3.4.1.2 Example: Bernoulli data model. We now provide an example of the proposed graphical elicitation procedure as it applies to a Bernoulli data model.

Suppose the Bernoulli model is chosen, and the expert selects a typical number of successes of $x = 50$ for a hypothetical future dataset of size $N = 100$, which results in an initial current step of $p^{(0)} = x/N = 0.5$. A proposed proportion p^{prop} is then drawn from a $\mathcal{N}(p^{(0)}, 0.05)$ proposal distribution.⁴ Next, a sample \mathbf{u} of size N is randomly drawn from the $\text{Unif}(0, 1)$ distribution and inverse transformed using the Bernoulli quantile function $F_p^{-1}(u)$ with both the current and proposed proportions to obtain

$$\begin{aligned}\mathbf{x}^{\text{current}} &= F_{p^{(0)}}^{-1}(\mathbf{u}), \text{ and} \\ \mathbf{x}^{\text{proposed}} &= F_{p^{\text{prop}}}^{-1}(\mathbf{u}).\end{aligned}$$

Bar charts of $\mathbf{x}^{\text{current}}$ and $\mathbf{x}^{\text{proposed}}$ are then displayed side-by-side, and the expert selects one of the five options. Recall these options consist of the following: (1) the proposed graph is more likely, (2) the current and proposed graphs are equally likely, (3) the current graph is three times more likely than the proposed, (4) the current graph is 25 times more likely than the proposed, or (5) the current graph is far more likely than the proposed. If the expert selects either option (1) or (2), the proportion p^{prop} associated with the proposed graph is automatically accepted and becomes the new current step of the sampler. If the expert selects one of the three odds options,

³ For the priors considered in this dissertation, the MLEs will always exist.

⁴ We found a proposal standard deviation of 0.05 works well for the Bernoulli data model.

input : N – the number of observations in a hypothetical future dataset
 x – a common measurement value for the hypothetical future dataset
(e.g. 50 successes for the binomial; count of 20 for the Poisson)
 f_θ – the data model family parameterized by $\theta \in \Theta$
 F_θ^{-1} – the inverse CDF for f_θ
 π_ϕ – the posterior model family parameterized by $\phi \in \mathbf{H}$
 σ_θ – the proposal standard deviation for θ
 \mathbf{O} – the odds for how likely the current plot is relative to the
proposed plot (e.g. $[1, 3, 25, 10^6]$)
 M – the total number of selections made (e.g. 100)

output: η^* , the hyperparameters for the prior on θ

```

1  $j \leftarrow 0$ 
2  $\theta^{(j)} \leftarrow$  the  $\theta$  corresponding to the  $x$  input by the expert
3 repeat
4   Sample  $\mathbf{u} = [u_1 \cdots u_N] \stackrel{iid}{\sim} \text{Unif}(0, 1)$ 
5   Sample  $\theta^{\text{prop}} \sim \mathcal{N}(\theta^{(j)}, \sigma_\theta)$ 
6    $\mathbf{x}^{\text{current}} = F_{\theta^{(j)}}^{-1}(\mathbf{v})$ 
7    $\mathbf{x}^{\text{proposed}} = F_{\theta^{\text{prop}}}^{-1}(\mathbf{v})$ 
8   Construct adjacent graphics with  $\mathbf{x}^{\text{current}}$  and  $\mathbf{x}^{\text{proposed}}$ 
9    $O^* \leftarrow$  the  $O_i$  corresponding to the expert's selection
10   $\alpha \leftarrow \min(1, 1/O^*)$ 
11  Sample  $v \sim \text{Unif}(0, 1)$ 
12  if  $v < \alpha$  then  $\theta^{(j+1)} \leftarrow \theta^{\text{prop}}$ 
13  else  $\theta^{(j+1)} \leftarrow \theta^{(j)}$ 
14   $j \leftarrow j + 1$ 
15 until  $j = M$ 
16  $\boldsymbol{\theta}^* \leftarrow [\theta^{(1)} \cdots \theta^{(M)}]$ 
17  $\eta^* \leftarrow \underset{\phi \in \mathbf{H}}{\text{argmax}} \pi_\phi(\boldsymbol{\theta}^*)$ 

```

Algorithm 3. The graphical prior elicitation procedure for data models with one unknown parameter.

the proposed value is accepted with only probability $1/O_i^{\text{selected}}$. If p^{prop} is rejected, $p^{(0)}$ remains the current step of the sampler. This process is repeated until the expert makes M selections. Finally, a Beta(α, β) prior is fit to the chain of parameter values from the sampler using maximum likelihood, resulting in the quantities $\hat{\alpha}_{\text{MLE}}$ and $\hat{\beta}_{\text{MLE}}$.

3.4.2 $\mathcal{N}(\mu, \sigma^2)$ Data Model with μ and σ^2 Unknown

The procedure as previously described naturally generalizes to any uniparameter data model. We now discuss the procedure for a multiparameter data model, the $\mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown.

For data models with only one unknown parameter θ , the expert made M total selections pertaining to θ . The $\mathcal{N}(\mu, \sigma^2)$ data model with σ^2 unknown has two unknown parameters $\boldsymbol{\theta} = [\mu, \sigma^2]$ and the expert must make M selections for each. We chose the conjugate prior for this model, a Normal-inverse-gamma($\mu_0, \lambda, \alpha, \beta$) prior on $\boldsymbol{\theta}$, as is common practice and as was done for the models discussed previously. The MLE $\hat{\boldsymbol{\eta}}_{\text{MLE}} = [\hat{\mu}_{0\text{MLE}}, \hat{\lambda}_{\text{MLE}}, \hat{\alpha}_{\text{MLE}}, \hat{\beta}_{\text{MLE}}]$ of the hyperparameters $\boldsymbol{\eta} = [\mu_0, \lambda, \alpha, \beta]$ is then computed once the expert has completed their selections.

The first step of the procedure entails asking the expert for two quantities: a typical measurement value x for a hypothetical future dataset of size N , and the largest possible value x_u for the dataset. Again, these values are only intended to calibrate the algorithm; they need not be correct in any conventional sense, and their precision is irrelevant. We then obtain initial values for the sampler as follows:

$$\begin{aligned}\theta_1^{(0)} &\stackrel{\text{set}}{=} x, \text{ and} \\ \theta_2^{(0)} &\stackrel{\text{set}}{=} \frac{x_u - x}{3},\end{aligned}$$

the latter of which is used as a rough application of the empirical rule.

A proposed mean θ_1^{prop} is then drawn from a $\mathcal{N}(\theta_1^{(0)}, \sigma_{\theta_1})$ proposal distribution. Next, a sample \mathbf{u} of size N is randomly drawn from a Unif(0,1) distribution and inverse-transformed using the Normal quantile function with means of $\theta_1^{(0)}$ and θ_1^{prop} and a common variance of $\theta_2^{(0)}$:

$$\begin{aligned}\mathbf{x}^{\text{current}} &= F_{\theta_1^{(0)}, \theta_2^{(0)}}^{-1}(\mathbf{u}), \text{ and} \\ \mathbf{x}^{\text{proposed}} &= F_{\theta_1^{\text{prop}}, \theta_2^{(0)}}^{-1}(\mathbf{u}).\end{aligned}$$

Histograms of $\mathbf{x}^{\text{current}}$ and $\mathbf{x}^{\text{proposed}}$ are then displayed side-by-side, and the expert must select one of five options, the same as those presented for the other data models.

If the expert selects an option where the proposed and current plots are equally likely or where the proposed plot is more likely, then θ_1^{prop} becomes the new current mean step $\theta_1^{(1)}$ in the sampler. On the other hand, if the expert selects one of the three options corresponding to the current graph being more likely than the proposed, then it is accepted with probability $1/O_i^{\text{selected}}$. If θ_1^{prop} is rejected, $\theta_1^{(1)} = \theta_1^{(0)}$.

After one proposal in the mean parameter, the algorithm next turns its attention to the variance parameter, with the mean fixed at $\theta_1^{(1)}$. A proposed variance θ_2^{prop} is drawn from a $\mathcal{N}(\theta_2^{(0)}, \sigma_{\theta_2})$ proposal distribution, and the previous process is repeated:

$$\begin{aligned}\mathbf{x}^{\text{current}} &= F_{\theta_1^{(1)}, \theta_2^{(0)}}^{-1}(\mathbf{u}), \text{ and} \\ \mathbf{x}^{\text{proposed}} &= F_{\theta_1^{(1)}, \theta_2^{\text{prop}}}^{-1}(\mathbf{u}).\end{aligned}$$

Histograms of $\mathbf{x}^{\text{current}}$ and $\mathbf{x}^{\text{proposed}}$ are again displayed side-by-side; the expert selects one of the five options; and $\theta_2^{(1)}$ is determined based on the option selected.

The algorithm then refocuses on the mean. Proposed and current samples are generated; histograms are plotted side-by-side; and the expert is tasked with selecting the best of the five options for how likely the two histograms are relative to one another. The procedure continues to alternate between the mean and variance, allowing only one parameter to vary at a time while fixing the other at its current value. This kind of movement has one key advantage: it allows the expert to make judgments according to one parameter at a time. If the variables were both proposed simultaneously, one might envision a scenario where an expert likes one graphic's location more but another one's scale more, and then cannot decide which to select.

The algorithm terminates after the expert has made M selections for each parameter, at which point hyperparameters are found by computing the MLE of the joint prior, resulting in a Normal-inverse-gamma $(\hat{\mu}_{0_{\text{MLE}}}, \hat{\lambda}_{\text{MLE}}, \hat{\alpha}_{\text{MLE}}, \hat{\beta}_{\text{MLE}})$ prior on $[\mu, \sigma^2]$. The full procedure is presented algorithmically in Algorithm 4.

input : N – the number of observations in a hypothetical future dataset
 x – a common measurement value for the hypothetical future dataset
 x_u – the largest possible value for the hypothetical future dataset
 f_θ – the Normal data model parameterized by $\theta = [\mu, \sigma^2] \in \Theta$
 F_θ^{-1} – the Normal inverse CDF parameterized by θ
 p_η – the Normal-inverse-gamma prior parameterized by
 $\eta = [\mu_0, \lambda, \alpha, \beta] \in \mathbf{H}$
 π_ϕ – the Normal-inverse-gamma posterior parameterized by $\phi \in \mathbf{H}$
 σ_θ – the proposal standard deviation for $\theta \in \Theta$
 \mathbf{O} – the odds for how likely the current plot is relative to the proposed plot (e.g. $[1, 3, 25, 10^6]$)
 M – the total number of selections made for each θ (e.g. 100)

output: η^* , the hyperparameters for the prior on θ

```

1  $j \leftarrow 0$ 
2  $\theta_1^{(j)} \leftarrow$  the  $\theta_1$  corresponding to the  $x$  input by the expert
3  $\theta_2^{(j)} \leftarrow (x_u - x)/3$ 
4 repeat
5   for  $i \leftarrow 1$  to 2 do
6     Sample  $\mathbf{u} = [u_1 \cdots u_N] \stackrel{iid}{\sim} \text{Unif}(0, 1)$ 
7     Sample  $\theta_i^{\text{prop}} \sim \mathcal{N}(\theta_i^{(j)}, \sigma_{\theta_i})$ 
8     if  $i = 1$  then  $\mathbf{x}^{\text{current}} = F_{\theta_i^{(j)}, \theta_{i+1}^{(j)}}^{-1}(\mathbf{v})$  and  $\mathbf{x}^{\text{proposed}} = F_{\theta_i^{\text{prop}}, \theta_{i+1}^{(j)}}^{-1}(\mathbf{v})$ 
9     else  $\mathbf{x}^{\text{current}} = F_{\theta_{i-1}^{(j+1)}, \theta_i^{(j)}}^{-1}(\mathbf{v})$  and  $\mathbf{x}^{\text{proposed}} = F_{\theta_{i-1}^{(j+1)}, \theta_i^{\text{prop}}}^{-1}(\mathbf{v})$ 
10    Construct adjacent graphics with  $\mathbf{x}^{\text{current}}$  and  $\mathbf{x}^{\text{proposed}}$ 
11     $O^* \leftarrow$  the  $O_i$  corresponding to the expert's selection
12     $\alpha \leftarrow \min(1, 1/O^*)$ 
13    Sample  $v \sim \text{Unif}(0, 1)$ 
14    if  $v < \alpha$  then  $\theta_i^{(j+1)} \leftarrow \theta_i^{\text{prop}}$ 
15    else  $\theta_i^{(j+1)} \leftarrow \theta_i^{(j)}$ 
16  end
17   $j \leftarrow j + 1$ 
18 until  $j = M$ 
19  $\Theta^* \leftarrow \begin{bmatrix} \theta_1^{(1)} & \theta_2^{(1)} \\ \vdots & \vdots \\ \theta_1^{(M)} & \theta_2^{(M)} \end{bmatrix}$ 
20  $\eta^* \leftarrow \underset{\phi \in \mathbf{H}}{\text{argmax}} \pi_\phi(\Theta^*)$ 

```

Algorithm 4. The graphical prior elicitation procedure for a $\mathcal{N}(\mu, \sigma^2)$ data model with both μ and σ^2 unknown.

3.5 Shiny Application

A free Shiny application developed for the four data models discussed in this chapter can be found at ccasement.shinyapps.io/graphicalElicitationMCMC (Chang et al., 2016). Its source code is on GitHub, a development platform that enables web-based version controlling of files and which is a hub for open-source code, at github.com/ccasement/graphicalElicitationMCMC. In this section we demonstrate the elicitation of a beta prior for a Bernoulli proportion through a series of screenshots of the app.

Suppose the expert expects $x = 67$ successes out of a hypothetical future dataset of size $N = 100$. They then proceed to the Rorschach training stage, where they are able to inspect randomly-generated datasets from Bernoulli distributions. Figure 3.5 displays bar charts of nine datasets of size $N = 100$ randomly generated from a $\text{Bernoulli}(p = 0.67)$ distribution, where the proportion of 0.67 corresponds to the $x/N = 67/100$ input by the expert. The expert is also able to go through Rorschach training for other Bernoulli distributions and can generate new sets of random samples for further training.

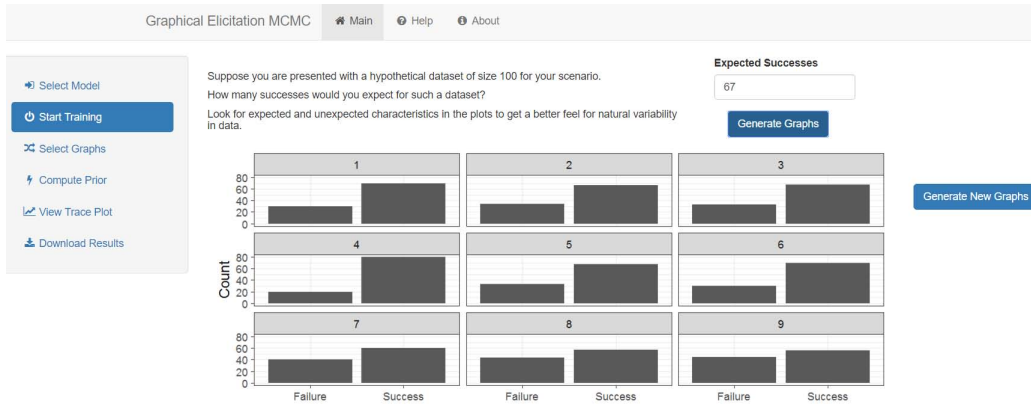


Figure 3.5. Before the graphical selection process begins, the expert can go through Rorschach training. Each bar chart displays a random sample of size $n = 100$ from the $\text{Bernoulli}(p = 0.67)$ distribution.

After finishing the Rorschach training the expert moves to the graphical elicitation procedure detailed in Section 3.4. Two bar charts – one for a current parameter value and another for a proposed parameter value – are presented to the expert, as displayed in Figure 3.6. The expert selects one of the five buttons, and new current and proposed datasets are generated and plotted.

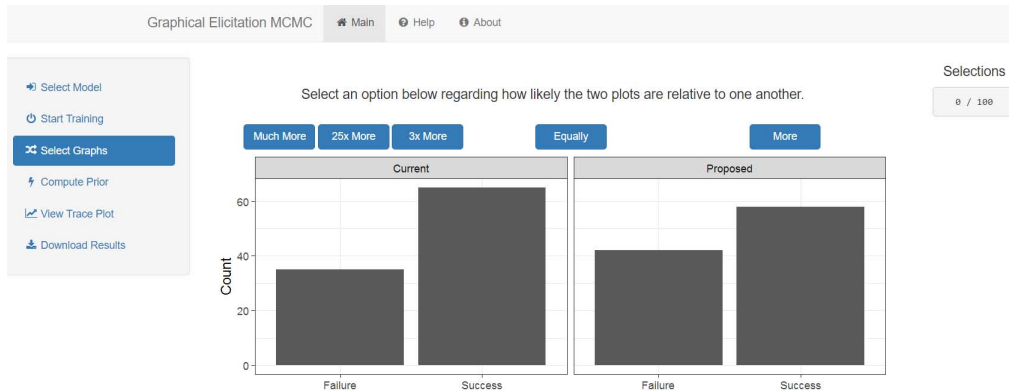


Figure 3.6. At each step of the selection process two graphics are presented to the expert, along with options for choosing between them.

The selections process concludes and a prior is computed once the expert has made 100 selections. Figure 3.7 displays information about the prior that is provided to the expert: the elicited prior family and the estimated hyperparameters, summaries of the prior, and a density plot. These plots and summaries enable the facilitator and expert to assess the adequacy of the prior distribution. In fact, additional options are provided to aid in the assessment, including one that allows users to view a kernel density estimate of the selections and another that calculates the probability the proportion is between any two values they specify. The application also provides a trace plot of the chain, as shown in Figure 3.8. Further, the expert can download a PDF report of the results as well as a CSV file containing the chain.

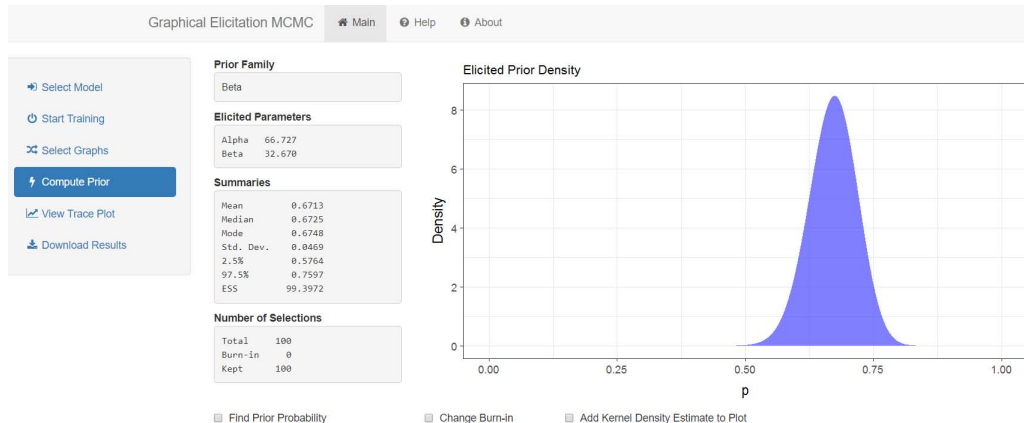


Figure 3.7. After the expert finishes the selection process, information regarding the elicited prior is presented.

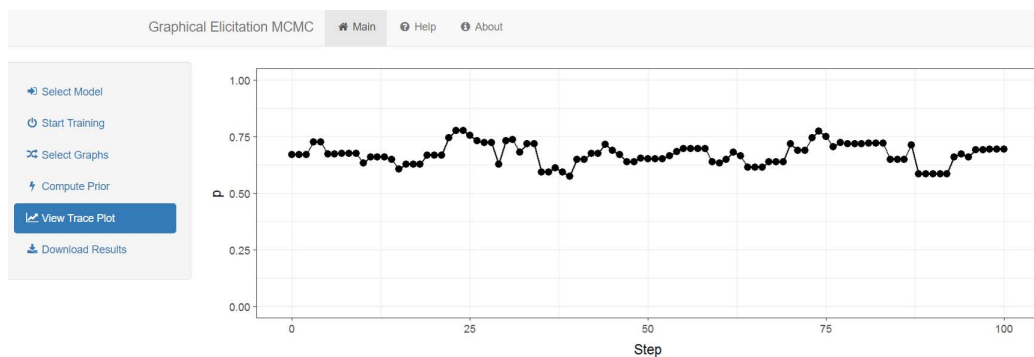


Figure 3.8. A trace plot of the accepted proportions is also available.

3.6 Technical Considerations

In the previous sections we laid out an intuitively motivated framework to think about a graphical solution to the problem of elicitation based on the same intuition as Markov chain Monte Carlo. We also provided a demonstration of an implementation of the algorithm. While the method has proved reasonable in basic scenarios, and the simulations in Chapter 4 support the basic efficacy of the strategy, the best demonstration of the soundness of the method lies in the elucidation of its parts. In many ways, this is the most we can hope for, as all elicitation strategies have a conceptual threshold beyond which no strategy can reach and assumptions must be

made. In this section we offer a technical framework in which to consider the problem, our proposed strategy of it, and the assumption it demands.

We begin with a description of the statistical components of the problem. We envision the expert as having obtained their current belief through both a theoretical framework, which is hard to model, and a previous dataset \mathbf{x} , considered to be a simple random sample from a data model $f(\mathbf{x}|\theta)$ assumed known up to θ . We suppose they first represented their ignorance about θ with an uninformative prior before seeing \mathbf{x} . Their ignorance of θ was then properly updated using Bayes' theorem after having seen \mathbf{x} to arrive at a posterior $f(\theta|\mathbf{x})$ that represents their current state of belief, i.e., the prior we are trying to elicit. We hope to learn this prior through a series of graphical selections the expert makes based on hypothetical future samples \mathbf{y} that are assumed to be simple random samples from the same data model $f(\mathbf{x}|\theta)$ with the same true unknown value of the parameter θ .

Once a graphic is selected (e.g. histogram), for a given sample size n of the hypothetical future sample \mathbf{y} , the assumed data model $f(\mathbf{x}|\theta)$ will induce a probability distribution on the space of graphics possible. Ultimately, the assumption is that the expert can reasonably differentiate probabilities on this scale. We now turn to the specifics.

The MCMC elicitation method is strongly based on Metropolis ideas, however, a few differences exist: (1) how the acceptance probabilities α are determined, (2) the possible set of values for α at any step of the sampler, and (3) the stationary distribution that results from the process. We now examine these differences.

Since we use a symmetric proposal distribution, the algorithm is a classic Metropolis algorithm. To correctly sample from the expert's prior, the algorithm requires the transition probability

$$\alpha = \min\left(1, \frac{\pi(\theta^{\text{prop}}|\mathbf{x})}{\pi(\theta^{(t)}|\mathbf{x})}\right). \quad (3.1)$$

For such a sampler, α can take on any value in the interval $(0, 1]$. The elicitation method, on the other hand, estimates the ratio of the target densities in (3.1) using the expert-selected ratio of probabilities of the proposed and current graphics based on \mathbf{x} :

$$\alpha = \min \left(1, \frac{P(G^{\text{prop}}|\mathbf{x})}{P(G^{(t)}|\mathbf{x})} \right) \approx \min \left(1, \frac{\pi(\theta^{\text{prop}}|\mathbf{x})}{\pi(\theta^{(t)}|\mathbf{x})} \right). \quad (3.2)$$

Since it is unreasonable to ask the expert to specify α on the entire interval, and in order to simplify the app, we round these probabilities to values in the set $\{10^{-6}, 1/25, 1/3, 1\}$ and present them as odds. The first three probabilities correspond to the current graphic being 10^6 times more likely, 25 times more likely, and three times more likely, respectively, than the proposed graphic, while an acceptance probability of one corresponds to the proposed graphic being equally like or more likely than the current graphic.

Because the set of possible acceptance probabilities used in the elicitation method differs from that used in Metropolis, we must examine the impact on the resulting distribution. We thus turn to properties of Markov chains. In order to ensure a Markov chain converges to a target distribution, certain criteria must be met. The chain must be (1) irreducible, (2) aperiodic, and (3) non-transient, and (4) the stationary distribution must be the same as the target distribution (Gelman et al., 2014). The MCMC elicitation method uses a continuous proposal distribution that allows any state to be reached from any other state, satisfying the irreducibility condition. The Markov chain is also aperiodic and non-transient, as the method employs a random walk. The distribution resulting from the elicitation method is thus stationary. While it is not the same as the target distribution, the true marginal posterior, the simulations in Chapter 4 demonstrate that the stationary distribution of the Markov chain is acceptably and quantifiably correct if the expert can specify the ratio of graphical probabilities presented above.

3.7 Conclusion

While methods and tools have been developed that enable experts to inject their beliefs into a Bayesian analysis through the use of an elicited prior, the methods possess certain drawbacks. To this end, we have proposed an interactive graphical procedure for prior elicitation that allows experts to work with data rather than parameters, as we believe experts can more reliably attest to the former than the latter. The free Shiny implementation not only enables experts to use the proposed methods, but it also leads them through the full elicitation process in a synergistic way, emphasizing often undervalued stages such as expert training and post-elicitation verification of the elicited prior, rather than focusing solely on the elicitation and fitting stages. Additionally, while the application allows experts to work with Bernoulli, Poisson, and Normal data models, all with conjugate priors, the procedure is capable of being extended to other data models and prior structures.

The MCMC elicitation scheme presented in this chapter opens the door to a new line of research for eliciting priors. Additional considerations include developing similar procedures for multivariate data models and theoretically connecting the two acceptance probabilities in (3.2).

CHAPTER FOUR

The Rigid Metropolis Algorithm for Approximate Bayesian Computation

Abstract

Posterior distributions play a fundamental role in the Bayesian paradigm, as they enable statistical inferences to be made. Advances in computational statistics combined with advances in computing over the past several decades have enabled those employing Bayesian methods to explore increasingly-complicated models. One computational method for understanding posteriors, the Metropolis-Hastings algorithm, has become popular for various reasons, including its abilities to (1) serve as a general purpose tool that typically does not depend heavily on the distribution sampled, (2) get by with only the un-normalized density $f(x)$, and (3) work in practice by providing sensible solutions to real world problems. In fact, the frequency with which researchers are employing Bayesian methods across the disciplines has increased as a direct result.

While Metropolis-Hastings presents clear strengths, however, it cannot be applied in all situations where an MCMC procedure is desired, such as when a discrete set of transition probabilities is demanded for the Markov chain. To this end, in this chapter we investigate the properties of a variation on Metropolis-Hastings that utilizes discrete sets of transition probabilities. We then apply the process to recently-developed graphical prior elicitation procedures.

4.1 Introduction

Over the past several decades the basic strategy underlying Monte Carlo methods, namely transformations of more basic randomly generated numbers, has been fortified with a rich collection of techniques from stochastic processes, especially dis-

crete time Markov chains. The underlying goal of all of the methods is the same: given a distribution $f(x)$, simulate independent and identically distributed (iid) draws from $f(x)$.

This transition has been perhaps most important in applications of Bayesian data analysis, where the target distribution $f(x)$ is the posterior distribution on the parameter of interest, given the data which is known up to a constant of proportionality. Particular instantiations of Monte Carlo methods, combined with the advent of the personal computer, have enabled Bayesian data analysis in ways previously impossible, bringing it to more than merely a viable alternative to frequentist methods. Many of these Monte Carlo methods, such as Metropolis, Metropolis-Hastings, and Gibbs sampling, among others, are built on a Markov chain Monte Carlo (MCMC) foundation.

Yet while these samplers – especially those with an MCMC framework – are widely used, they cannot be employed for every situation that calls for them. For instance, transition probabilities for such samplers can often take on any value in the interval $(0, 1]$.¹ However, if a discrete set of $r \in \mathbb{Z}^+$ transition probabilities is deemed necessary, these samplers, as they stand, cannot be utilized. To address this drawback, we propose a variation on the Metropolis algorithm that disallows the full continuum of values $(0, 1]$ as transition probabilities. We refer to this method as “rigid Metropolis.” In this chapter we construct a framework to quantitatively assess the asymptotic distribution of the rigid Metropolis algorithm in comparison to its standard Metropolis counterpart using the total variation distance and Monte Carlo simulation. We then use this framework to select near-optimal rigid sets of various sizes.

The chapter proceeds as follows. In Section 4.2 we discuss the total variation distance (TVD) metric and describe how it can be used to assess the relationship

¹ Gibbs sampling is an example of a method where the transition probability does not take on any value in this interval. Rather, it is always one.

between standard Metropolis and rigid Metropolis. In Section 4.3 we formally introduce the new variation on Metropolis, rigid Metropolis, that lends support to the transitions made in the graphical elicitation method proposed in Chapter 3. We then use the TVD to assess the accuracy of posteriors found via applications of the rigid MCMC algorithm and formulate a process for obtaining near-optimal rigid sets of a given size. We next present simulation results for three data models discussed in this dissertation – Bernoulli(p), Poisson(λ), and $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known – and discuss their impacts on the proposed elicitation method. We conclude with a summary in Section 4.4.

4.2 Total Variation Distance

Akin to formulations such as the Kolmogorov-Smirnov distance and the Kullback-Leibler divergence, the total variation distance is a classic metric on the space of probability measures defined on the same measurable space (Ω, \mathcal{B}) . It is defined as

$$\delta(P, Q) = \sup_{A \in \mathcal{B}} |P(A) - Q(A)| \quad (4.1)$$

for probability measures P and Q on the measurable space (Ω, \mathcal{B}) .

Remarkably, it is also known that, for probability distributions on \mathbb{R} , the TVD $\delta(P, Q)$ takes the form

$$\delta(P, Q) = \frac{1}{2} \int_{\mathbb{R}} |p(x) - q(x)| dx. \quad (4.2)$$

For clarity, we provide a proof of this result below. The proof is a modified version of the one presented in Resnick (2013).

Proposition 4.1 Suppose P and Q are probability distributions with densities p and q with respect to Lebesgue measure on $(\mathbb{R}, \mathcal{B})$, so that, for example, if $A \in \mathcal{B}$, $P(A) = \int_A p(x) dx$. Then

$$\sup_{A \in \mathcal{B}} |P(A) - Q(A)| = \frac{1}{2} \int_{\mathbb{R}} |p(x) - q(x)| dx.$$

Proof. First, note that

$$\int_{\mathbb{R}} (p(x) - q(x)) dx = \int_{\mathbb{R}} p(x) dx - \int_{\mathbb{R}} q(x) dx = 1 - 1 = 0.$$

Consequently, for any event $A \in \mathcal{B}$,

$$\int_A (p(x) - q(x)) dx + \int_{A^c} (p(x) - q(x)) dx = 0.$$

It follows that

$$\left| \int_A (p(x) - q(x)) dx \right| = \left| \int_{A^c} (p(x) - q(x)) dx \right|.$$

Thus,

$$\begin{aligned} 2|P(A) - Q(A)| &= 2 \left| \int_A (p(x) - q(x)) dx \right| \\ &= \left| \int_A (p(x) - q(x)) dx \right| + \left| \int_{A^c} (p(x) - q(x)) dx \right| \\ &\leq \int_A |p(x) - q(x)| dx + \int_{A^c} |p(x) - q(x)| dx \\ &= \int_{\mathbb{R}} |p(x) - q(x)| dx. \end{aligned} \tag{4.3}$$

Consequently,

$$|P(A) - Q(A)| \leq \frac{1}{2} \int_{\mathbb{R}} |p(x) - q(x)| dx.$$

Since the above inequality holds for all $A \in \mathcal{B}$,

$$\sup_{A \in \mathcal{B}} |P(A) - Q(A)| \leq \frac{1}{2} \int_{\mathbb{R}} |p(x) - q(x)| dx. \tag{4.4}$$

Now, consider the set $A = \{x \in \mathbb{R} : p(x) \geq q(x)\}$. $A \in \mathcal{B}$ is measurable since both p and q are measurable functions (w.r.t. the Lebesgue measure on \mathbb{R}). From (4.3) we know

$$2|P(A) - Q(A)| = \left| \int_A (p(x) - q(x)) dx \right| + \left| \int_{A^c} (p(x) - q(x)) dx \right|.$$

It follows from (4.3) that

$$2|P(A) - Q(A)| = \int_A |p(x) - q(x)| dx + \int_{A^c} |p(x) - q(x)| dx,$$

equality that results from the nonnegative and nonpositive integrands, respectively, by the choice of A . Of course,

$$\int_A |p(x) - q(x)| dx + \int_{A^c} |p(x) - q(x)| dx = \int_{\mathbb{R}} |p(x) - q(x)| dx$$

so that, for this carefully selected A ,

$$|P(A) - Q(A)| = \frac{1}{2} \int_{\mathbb{R}} |p(x) - q(x)| dx.$$

Thus, the bound determined in (4.4) is always achieved by some A , confirming the proposition. □

In the next section we present the rigid Metropolis algorithm and use the TVD to assess the resulting posteriors. The PDFs for the three posterior distributions of interest – beta, gamma, and Normal – are nice, so that the integral in Proposition 4.1 can be computed efficiently numerically with a high degree of precision.

4.3 Rigid MCMCs

We now turn to a more formal description of rigid Metropolis and consider its resulting asymptotic distributions in relation to the asymptotic distributions of standard Metropolis in light of the TVD.

4.3.1 General Scheme

The underlying process behind rigid Metropolis mimics that of standard Metropolis at all stages except for the acceptance probability stage. In a standard Metropolis-Hastings sampler, the acceptance probability $\alpha \in (0, 1]$ of a proposed step θ^{prop} is defined as

$$\alpha = \min \left(1, \frac{g(\theta^{(t)} | \theta^{\text{prop}})}{g(\theta^{\text{prop}} | \theta^{(t)})} \frac{f(\theta^{\text{prop}})}{f(\theta^{(t)})} \right), \quad (4.5)$$

for a target density $f(\cdot)$ and proposal density $g(\cdot|\cdot)$. In fact, when a symmetric proposal is chosen for $g(\cdot|\cdot)$, (4.5) simplifies to

$$\alpha = \min \left(1, \frac{f(\theta^{\text{prop}})}{f(\theta^{(t)})} \right).$$

Once α is computed, there is a $(100\alpha)\%$ chance θ^{prop} is accepted.

With rigid Metropolis, on the other hand, once α is calculated, it is rounded to the nearest probability p_{\star}^{rigid} in a pre-specified vector of r probabilities $\mathcal{P} = \{p_1^{\text{rigid}}, \dots, p_r^{\text{rigid}}\}$ that we call the rigid set. The acceptance probability of the rigid Metropolis algorithm is thus $\alpha^* = \min(1, p_{\star}^{\text{rigid}})$, resulting in a $(100\alpha^*)\%$ chance the proposed step is accepted. The remainder of the algorithm also follows that of Metropolis. The full rigid Metropolis process is detailed in algorithmic format in Algorithm 5 for data models with one unknown parameter, though, similar to standard Metropolis, it generalizes to models with p unknowns. We note in passing that rigid Metropolis can be generalized to Metropolis-Hastings, but we do not consider it further in this work.

4.3.2 Applications to Graphical Prior Elicitation

The stochastic procedure proposed in Chapter 3 enables experts to elicit prior distributions using a finite number of options of odds – in the form of buttons in the Shiny app – that represent the likelihood of the current plot relative to that of the proposed plot. Each of these odds is, in fact, the reciprocal of an acceptance probability used in the rigid set. There is thus a bijective mapping between the probabilities $p_i^{\text{rigid}} \in (0, 1]$ and odds $O_i^{\text{rigid}} \in \mathbb{R}^+$. This mapping is $h : (0, 1] \rightarrow \mathbb{R}^+$, where $O^{\text{rigid}} = h(p^{\text{rigid}}) = 1/p^{\text{rigid}}$.

Before a rigid Metropolis procedure is performed, the number of probabilities, their values, and appropriate proposal standard deviations for the sampler must be set. Ideally, for a given number of such probabilities they are determined in such a way that the asymptotic distribution of the rigid Metropolis procedure is as close as

<p>input : $\pi(\theta)$ – the posterior for θ given the missing previously-observed dataset \mathbf{x}</p> <p>$\theta^{(1)}$ – the initial value for θ (e.g. θ)</p> <p>σ_θ – the proposal standard deviation for θ</p> <p>\mathcal{P} – the rounded acceptance probabilities (e.g. $\{10^{-6}, 0.1, 1\}$)</p> <p>M – the number of MCMC iterations (e.g. 5,000)</p> <p>output: $\boldsymbol{\theta}$, samples from the rigid posterior of θ</p> <pre> 1 $j \leftarrow 1$ 2 repeat 3 Sample $\theta^{\text{prop}} \sim \mathcal{N}(\theta^{(j)}, \sigma_\theta)$ 4 $r \leftarrow \frac{\pi(\theta^{\text{prop}})}{\pi(\theta^{(j)})}$ 5 $p_\star^{\text{rigid}} \leftarrow$ the $p_i^{\text{rigid}} \in \mathcal{P}$ nearest r 6 $\alpha^\star \leftarrow \min(1, p_\star^{\text{rigid}})$ 7 Sample $u \sim \text{Unif}(0, 1)$ 8 if $u < \alpha^\star$ then $\theta^{(j+1)} \leftarrow \theta^{\text{prop}}$ 9 else $\theta^{(j+1)} \leftarrow \theta^{(j)}$ 10 $j \leftarrow j + 1$ 11 until $j = M$ 12 $\boldsymbol{\theta} \leftarrow [\theta^{(1)} \dots \theta^{(M)}]$ </pre>

Algorithm 5. General scheme of the rigid Metropolis process for data models with one unknown parameter.

possible to the asymptotic distribution of the standard Metropolis procedure. In this chapter we propose a strategy to obtain near-optimal rigid sets of lengths $r = 2, 3$, and 4, and for the proposal standard deviations σ_θ used in the graphical elicitation method in Chapter 3.

4.3.2.1 TVD for rigid Metropolis. When using MCMC procedures in a Bayesian analysis, the target distribution is typically the joint posterior distribution of the parameters. To assess the ability of rigid Metropolis to accurately approximate these distributions, we consider the TVD between the target posterior and that found using the rigid Metropolis procedure.

Suppose a data model f_θ is assumed with $\theta \in \Theta$ for a dataset \mathbf{x} , and an uninformative conjugate prior π_η with $\eta \in \mathbf{H}$ is assumed for θ . We first find the target posterior $\pi_{\text{target}}(\theta|\mathbf{x})$, the asymptotic distribution of a chain using standard Metropolis, which is known for all distributions in this chapter. We then find the posterior $\pi_{\text{rigid}}(\theta|\mathbf{x})$ using the rigid Metropolis process described in Section 4.3.1.

It is worth noting that in the beta case, for example, we assume that the prior is a Beta(1,1) so that the target posterior is also a beta (it is conjugate). Under standard Metropolis, that is also the stationary distribution of the chain. However, it is unproven as to whether the stationary distribution of the rigid Metropolis is, in fact, beta. In simulations it appeared to be so, and we consequently used its samples to fit a beta, but in general it need not be. This does not present a problem for the algorithm: we could have simply used the samples to determine an $\hat{f}(\theta)$ (a kernel density estimate) and then used $\hat{f}(\theta)$ in the integral formulation of the TVD. We chose not to for simplicity sake.

With the posteriors $\pi_{\text{target}}(\theta|\mathbf{x})$ and $\pi_{\text{rigid}}(\theta|\mathbf{x})$ in hand, we can compute the TVD between them:

$$\begin{aligned} \delta(\pi_{\text{target}}(\theta|\mathbf{x}), \pi_{\text{rigid}}(\theta|\mathbf{x})) &= \sup_{\theta \in \Theta} |\pi_{\text{target}}(\theta|\mathbf{x}) - \pi_{\text{rigid}}(\theta|\mathbf{x})| \\ &\approx \frac{1}{2} \int_{\Theta} |f_{\text{target}}(\theta|\mathbf{x}) - f_{\text{rigid}}(\theta|\mathbf{x})| d\theta, \end{aligned} \quad (4.6)$$

where $f_{\text{target}}(\theta|\mathbf{x})$ and $f_{\text{rigid}}(\theta|\mathbf{x})$ are the PDFs corresponding to the target and rigid posteriors. We elected to use the form of the TVD specified in (4.6) due to the ease with which it can be computed numerically between continuous distributions, a characteristic of all posteriors covered in this dissertation.

Working with rigid Metropolis, however, requires a pre-specified vector of probabilities $\mathcal{P} = \{p_1^{\text{rigid}}, \dots, p_r^{\text{rigid}}\}$ of length r . We now formulate a procedure for determining an optimal \mathcal{P} .

4.3.2.2 *Rigid Metropolis transition probabilities.* When determining appropriate rigid probabilities for $\mathcal{P} = \{p_1^{\text{rigid}}, \dots, p_r^{\text{rigid}}\}$, it seems obvious that r should be at least two, and those should indicate roughly “do not move” or “move”: $p_1^{\text{rigid}} = 10^{-6}$ and $p_r^{\text{rigid}} = 1$, for all lengths r . It is impossible to achieve an acceptance probability of exactly zero in the Metropolis algorithm, so we instead selected a value close to 0. This smallest probability, which corresponds to the current plot being $1/10^{-6} = 10^6$ times more likely than the proposed plot, virtually prevents the expert from getting stuck in an undesirable part of the parameter space. With the expert making a number of selections (e.g. 100) much smaller than the typical number of iterations for an MCMC sampler, this choice was deemed necessary to support the elicitation of a prior that represents the expert’s opinions. Next, the largest probability in \mathcal{P} was fixed at 1, as this corresponds to the proposed plot being either equally likely as, or more likely than, the current plot. Such is the case when comparing proposed and current parameters in standard Metropolis – if the proposed step is at least as equally likely as the current step, the proposed is accepted with probability one.

With the lowest and highest values in \mathcal{P} set, we now turn to finding additional probabilities for \mathcal{P} . For a fixed r and θ , our goal is to find the set of probabilities that minimizes the expected total variation distance between $\pi_{\text{target}}(\theta|\mathbf{x})$ and $\pi_{\text{rigid}}(\theta|\mathbf{x})$:

$$\mathcal{P}^*(\theta) = \underset{\mathcal{P}}{\operatorname{argmin}} E\left[\delta(\pi_{\text{target}}(\theta|\mathbf{x}), \pi_{\text{rigid}}(\theta|\mathbf{x})) | \theta, \mathcal{P}\right]. \quad (4.7)$$

Solving this optimization problem exactly for a given value of θ , however, is unnecessary for our purposes. The odds the expert must choose from in the Shiny app for the stochastic procedure must be ones they can reliably attest to. Thus, the probabilities considered for \mathcal{P} must result in such odds. Accounting for the expert’s ability to accurately select the buttons for the odds in the Shiny app as well as the computational complexity of the process discussed later in this section, while simultaneously ensuring the theoretical MCMC foundation of the process is maintained, we considered

$p_i^{\text{rigid}} \in \{0.02, 0.04, 0.06, \dots, 0.98\}$ as candidate probabilities for $p_2^{\text{rigid}}, \dots, p_{r-1}^{\text{rigid}} \in \mathcal{P}$, when $r > 2$.

To minimize (4.7) for a fixed rigid set of size $r \in \{2, 3, 4, \dots\}$, we do as follows. First randomly generate a sample of size n from f_θ , since the dataset \mathbf{x} previously seen by the expert is unknown. Second, find the target posterior $\pi_{\text{target}}(\theta|\mathbf{x})$ for θ ; this is a simple task since each of the priors used is conjugate. Next, approximate the posterior $\pi_{\text{rigid}}(\theta|\mathbf{x})$ for θ based on M iterations of the rigid Metropolis process outlined in Algorithm 5. This is done in a two-stage process: first generate the values with the rigid Metropolis algorithm, then use those to fit parameters values $\hat{\boldsymbol{\eta}}$ employing, for example, maximum likelihood. After finding $\pi_{\text{rigid}}(\theta|\mathbf{x})$, calculate the total variation distance $\delta(\pi_{\text{target}}(\theta|\mathbf{x}), \pi_{\text{rigid}}(\theta|\mathbf{x}))$ between the two posteriors. Perform this process T times and average the total variation distances for each combination of probabilities in \mathcal{P} . Then repeat this entire procedure for all c possible combinations of r candidate probabilities in the rigid set, resulting in c average total variation distances. With the ultimate goal being to minimize $E\left[\delta(\pi_{\text{target}}(\theta|\mathbf{x}), \pi_{\text{rigid}}(\theta|\mathbf{x}))|\theta, \mathcal{P}\right]$, select the \mathcal{P} that results in the smallest average total variation distance $\bar{\delta}(\pi_{\text{target}}(\theta|\mathbf{x}), \pi_{\text{rigid}}(\theta|\mathbf{x}))$. This grid-search style process is further detailed in Algorithm 6.

4.3.2.3 Rigid Metropolis proposal standard deviations. The choice of a proposal standard deviation σ_θ with data model f_θ is another important consideration when working with MCMC methods. If σ_θ is too small, then the sampler will not be very efficient; its values will be highly autocorrelated. As a consequence for our application, the expert will have difficulty distinguishing between the proposed and current plots, as the proposed and current parameter values will often be close in magnitude. Additionally, far too many selections will be demanded of the expert. On the other hand, if σ_θ is too large, many proposed values will be rejected, as they represent unrealistic scenarios or, in some cases, will be outside the parameter space.

input : f_θ – the data model parameterized by $\theta \in \Theta$
 $\pi(\theta)$ – the posterior for θ given the missing previously-observed dataset \mathbf{x}
 $\pi_{\text{target}}(\theta|\mathbf{x})$ – the target posterior
 $\pi_{\text{rigid}, \phi}(\theta|\mathbf{x})$ – the posterior resulting from the rigid Metropolis process and parameterized by ϕ
 $\theta^{(0)}$ – the initial value for θ (e.g. θ)
 σ_θ – the proposal standard deviation for θ
 \mathcal{P} – the rounded acceptance probabilities (e.g. $\{10^{-6}, 0.2, 1\}$)
 M – the number of MCMC iterations (e.g. 5,000)
 T – the number of iterations of the entire process (e.g. 1,000)

output: $\bar{\delta}(\pi_{\text{target}}(\theta|\mathbf{x}), \pi_{\text{rigid}, \phi}(\theta|\mathbf{x}))$, the average TVD between $\pi_{\text{target}}(\theta|\mathbf{x})$ and $\pi_{\text{rigid}, \phi}(\theta|\mathbf{x})$

```

1  $i \leftarrow 1$ 
2 repeat
3   Sample  $\mathbf{x} = [x_1 \cdots x_N] \stackrel{iid}{\sim} f_\theta$ 
4    $j \leftarrow 0$ 
5   repeat
6     Sample  $\theta^{\text{prop}} \sim \mathcal{N}(\theta^{(j)}, \sigma_\theta)$ 
7      $r \leftarrow \frac{\pi(\theta^{\text{prop}})}{\pi(\theta^{(j)})}$ 
8      $p_\star^{\text{rigid}} \leftarrow$  the  $p_i^{\text{rigid}} \in \mathcal{P}$  nearest  $r$ 
9      $\alpha^\star \leftarrow \min(1, p_\star^{\text{rigid}})$ 
10    Sample  $u \sim \text{Unif}(0, 1)$ 
11    if  $u < \alpha^\star$  then  $\theta^{(j+1)} \leftarrow \theta^{\text{prop}}$ 
12    else  $\theta^{(j+1)} \leftarrow \theta^{(j)}$ 
13     $j \leftarrow j + 1$ 
14  until  $j = M$ 
15   $\boldsymbol{\theta} \leftarrow [\theta^{(1)} \cdots \theta^{(M)}]$ 
16   $\phi_i^\star \leftarrow \underset{\phi \in \mathbf{H}}{\text{argmax}} \pi_{\text{rigid}, \phi}(\theta|\mathbf{x})$ 
17   $\delta_i \leftarrow \frac{1}{2} \int_{\Theta} |\pi_{\text{target}}(\theta|\mathbf{x}) - \pi_{\text{rigid}, \phi_i^\star}(\theta|\mathbf{x})| d\theta$ 
18   $i \leftarrow i + 1$ 
19 until  $i = T$ 
20  $\bar{\delta} \leftarrow \frac{1}{T} \sum_i \delta_i$ 

```

Algorithm 6. General scheme for assessing the accuracy of posteriors found using rigid Metropolis for data models with one unknown parameter.

As the expert makes a number of selections in the app that is small relative to the number of iterations typically used in practice, the proposal standard deviation used is an important choice for the successful elicitation of a prior that represents the expert’s beliefs. We found proposal standard deviations of $\sigma_p = 0.05$, $\sigma_\lambda = \sqrt{\lambda^{(0)}}$, and $\sigma_\mu = 2\sqrt{\sigma^2}$ to be reasonable for $\text{Bernoulli}(p)$, $\text{Poisson}(\lambda)$, and $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known data models, respectively. These standard deviations allow for efficient exploration of the parameter space when utilizing rigid Metropolis as well as in the graphical elicitation procedures, while being sufficiently large that they enable the expert to distinguish between plots in the Shiny app and reduce the need for thinning, maximizing the information gained from every selection.

We now assess the accuracy of the rigid Metropolis process for three data models in this dissertation: $\text{Bernoulli}(p)$, $\text{Poisson}(\lambda)$, and $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known. For each model we computed the average total variation distance between $\pi_{\text{target}}(\theta|\mathbf{x})$ and $\pi_{\text{rigid}}(\theta|\mathbf{x})$ for various distributions: Bernoulli with probabilities of 0.1 through 0.9, in increments of 0.1; Poisson with rates of 0, 5, 10, 25, 50, 100; and $\mathcal{N}(\mu, \sigma^2)$ with means of 0, 10, 25, 50, 100, and 500, all with a variance of 100.² Additionally, for each scenario above, we used acceptance probability vectors of lengths $r = 2, 3$, and 4. Further, for each case, we used $T = 1,000$ total iterations of $n = 100$ observations, each with $M = 5,000$ MCMC iterations.

4.3.2.4 Rigid Metropolis results for $r = 2$ rigid transition probabilities. We first examine the case where $r = 2$, for which the resulting vector of rigid MCMC transition probabilities is $\mathcal{P} = \{10^{-6}, 1\}$. Summaries of the resulting total variation distances between $\pi_{\text{target}}(\theta|\mathbf{x})$ and $\pi_{\text{rigid}}(\theta|\mathbf{x})$ for all cases considered can be found in Table 4.1. As the average total variation distances are large, two probabilities are not sufficient for \mathcal{P} . We therefore consider scenarios with $r = 3$ transition probabilities.

² Additional variances were considered, and the same conclusions as those discussed in the following sections can be drawn.

Table 4.1. Total variation distance summaries for Bernoulli(p), Poisson(λ), and $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on $\mathcal{P} = \{10^{-6}, 1\}$, indicating that merely rounding the transition probability in the Metropolis algorithm yields an unacceptably poor rigid approximation to standard Metropolis.

Data Model	Parameter	Mean δ	SD of δ	95% CI for δ
Bernoulli(p)	0.1	0.5381	0.0192	(0.5022, 0.5760)
	0.2	0.5194	0.0169	(0.4890, 0.5545)
	0.3	0.5096	0.0137	(0.4832, 0.5375)
	0.4	0.5035	0.0120	(0.4807, 0.5269)
	0.5	0.5011	0.0106	(0.4795, 0.5219)
	0.6	0.5031	0.0114	(0.4811, 0.5250)
	0.7	0.5086	0.0141	(0.4817, 0.5387)
	0.8	0.5198	0.0166	(0.4884, 0.5530)
	0.9	0.5389	0.0185	(0.5031, 0.5771)
Poisson(λ)	1	0.4070	0.0082	(0.3901, 0.4220)
	5	0.4054	0.0078	(0.3901, 0.4211)
	10	0.4055	0.0079	(0.3897, 0.4209)
	25	0.4057	0.0076	(0.3909, 0.4203)
	50	0.4056	0.0081	(0.3891, 0.4212)
	100	0.4053	0.0080	(0.3891, 0.4200)
$\mathcal{N}(\mu, \sigma^2 = 100)$	0	0.4068	0.0066	(0.3937, 0.4200)
	10	0.4065	0.0065	(0.3930, 0.4184)
	25	0.4069	0.0064	(0.3939, 0.4186)
	50	0.4066	0.0063	(0.3941, 0.4187)
	100	0.4065	0.0066	(0.3934, 0.4187)
	500	0.4064	0.0064	(0.3939, 0.4189)

4.3.2.5 Rigid Metropolis results for $r = 3$ rigid transition probabilities. For the case where $r = 3$, $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, 1\}$. The resulting average total variation distances for $p_2^{\text{rigid}} \in \{0.02, \dots, 0.98\}$ are plotted in Figures 4.1 through 4.3 for varying values of the Bernoulli proportion $\theta = p$, the Poisson rate $\theta = \lambda$, and the Normal mean $\theta = \mu$. For all three data models, we find that \mathcal{P} is stable across different values of the underlying value of θ . We also find the optimal (or near-optimal) rigid set to be $\mathcal{P} = \{10^{-6}, 0.1, 1\}$ for the Bernoulli data models and $\mathcal{P} = \{10^{-6}, 0.06, 1\}$ for the Poisson and Normal data models.

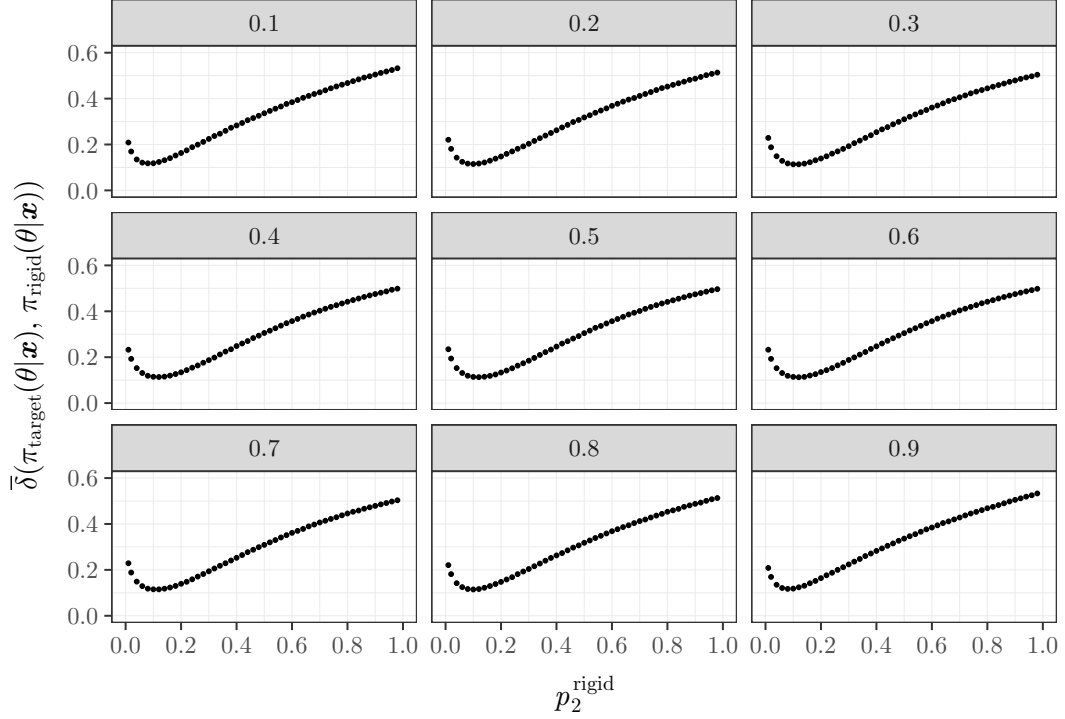


Figure 4.1. Average total variation distances for Bernoulli(p) data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, 1\}$. Plot labels communicate Bernoulli probabilities used.

Summaries of the total variation distances for $r = 3$ are presented in Table 4.2. While the average total variation distances have decreased substantially compared to those for $r = 2$, they still remain larger than desirable. Hence, three probabilities are insufficient for $\mathbf{p}^{\text{rigid}}$, and we next consider scenarios with $r = 4$ transition probabilities.

4.3.2.6 Rigid Metropolis results for $r = 4$ rigid transition probabilities. For the case where $r = 4$, $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, p_3^{\text{rigid}}, 1\}$. The resulting average total variation distances for all cases considered can be seen in Figures 4.4 through 4.6. For each distribution, the average total variation distance strongly depends on the value of p_2^{rigid} , whereas it does not strongly depend on the value of p_3^{rigid} . Based on the plots, the average distance appears smallest for $p_2^{\text{rigid}} \in (0.02, 0.12)$ for Bernoulli distributions and $p_2^{\text{rigid}} \in (0.02, 0.06)$ for Poisson and Normal distributions, and for many values

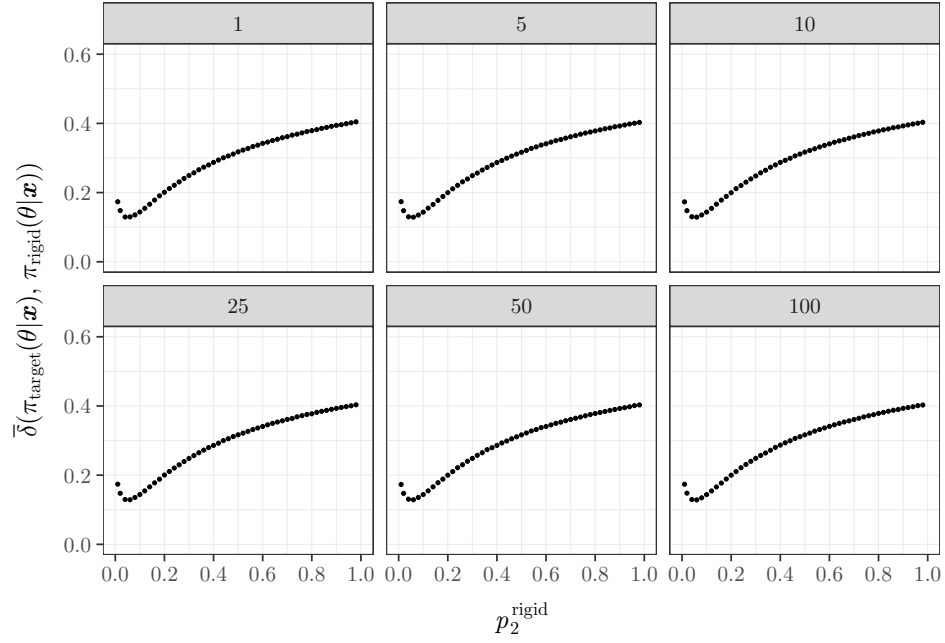


Figure 4.2. Average total variation distances for $\text{Poisson}(\lambda)$ data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, 1\}$. Plot labels communicate Poisson rates used.

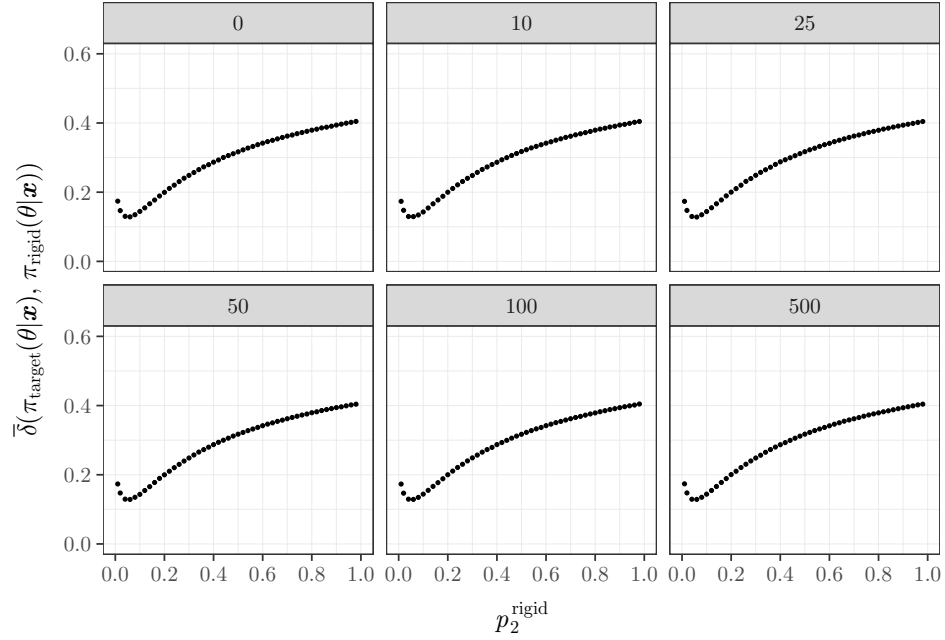


Figure 4.3. Average total variation distances for $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, 1\}$. Plot labels communicate Normal means used.

Table 4.2. Total variation distance summaries for Bernoulli(p), Poisson(λ), and $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, 1\}$, indicating that three rigid probabilities improve significantly on two (simple rounding), yet still leaving around a 10% worst-case error.

Data Model	Parameter	p_2^{rigid}	Mean δ	SD of δ	95% CI for δ
Bernoulli(p)	0.1	0.08	0.1180	0.0125	(0.0938, 0.1439)
	0.2	0.10	0.1150	0.0132	(0.0913, 0.1434)
	0.3	0.10	0.1137	0.0127	(0.0897, 0.1390)
	0.4	0.12	0.1137	0.0139	(0.0878, 0.1416)
	0.5	0.12	0.1130	0.0141	(0.0864, 0.1409)
	0.6	0.12	0.1127	0.0136	(0.0874, 0.1401)
	0.7	0.10	0.1152	0.0144	(0.0869, 0.1420)
	0.8	0.10	0.1147	0.0135	(0.0889, 0.1429)
	0.9	0.08	0.1172	0.0122	(0.0948, 0.1411)
Poisson(λ)	1	0.04	0.1295	0.0150	(0.0987, 0.1596)
	5	0.06	0.1290	0.0152	(0.1007, 0.1599)
	10	0.06	0.1291	0.0160	(0.0985, 0.1598)
	25	0.06	0.1290	0.0167	(0.0966, 0.1620)
	50	0.06	0.1289	0.0158	(0.0970, 0.1581)
	100	0.06	0.1285	0.0163	(0.0966, 0.1595)
$\mathcal{N}(\mu, 100)$	0	0.06	0.1287	0.0128	(0.1044, 0.1548)
	10	0.06	0.1294	0.0134	(0.1032, 0.1544)
	25	0.06	0.1281	0.0133	(0.1031, 0.1542)
	50	0.06	0.1287	0.0127	(0.1047, 0.1536)
	100	0.06	0.1287	0.0132	(0.1027, 0.1544)
	500	0.06	0.1288	0.0128	(0.1043, 0.1531)

of $p_3^{\text{rigid}} \in (p_2^{\text{rigid}}, 1)$. To simplify the options the expert must choose from in the Shiny app, we want p_3^{rigid} to be larger than p_2^{rigid} by an amount that presents a clear distinction between the odds $O_2^{\text{rigid}} = 1/p_2^{\text{rigid}}$ and $O_3^{\text{rigid}} = 1/p_3^{\text{rigid}}$. We determined $p_2^{\text{rigid}} = 0.04$ and $p_3^{\text{rigid}} = 0.34$, which correspond to $O_2^{\text{rigid}} = 25$ and $O_3^{\text{rigid}} \approx 3$, to reasonably satisfy these criteria.

The total variation distances for rigid MCMCs with $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$ and for each model considered are summarized in Table 4.3 and plotted in Fig-

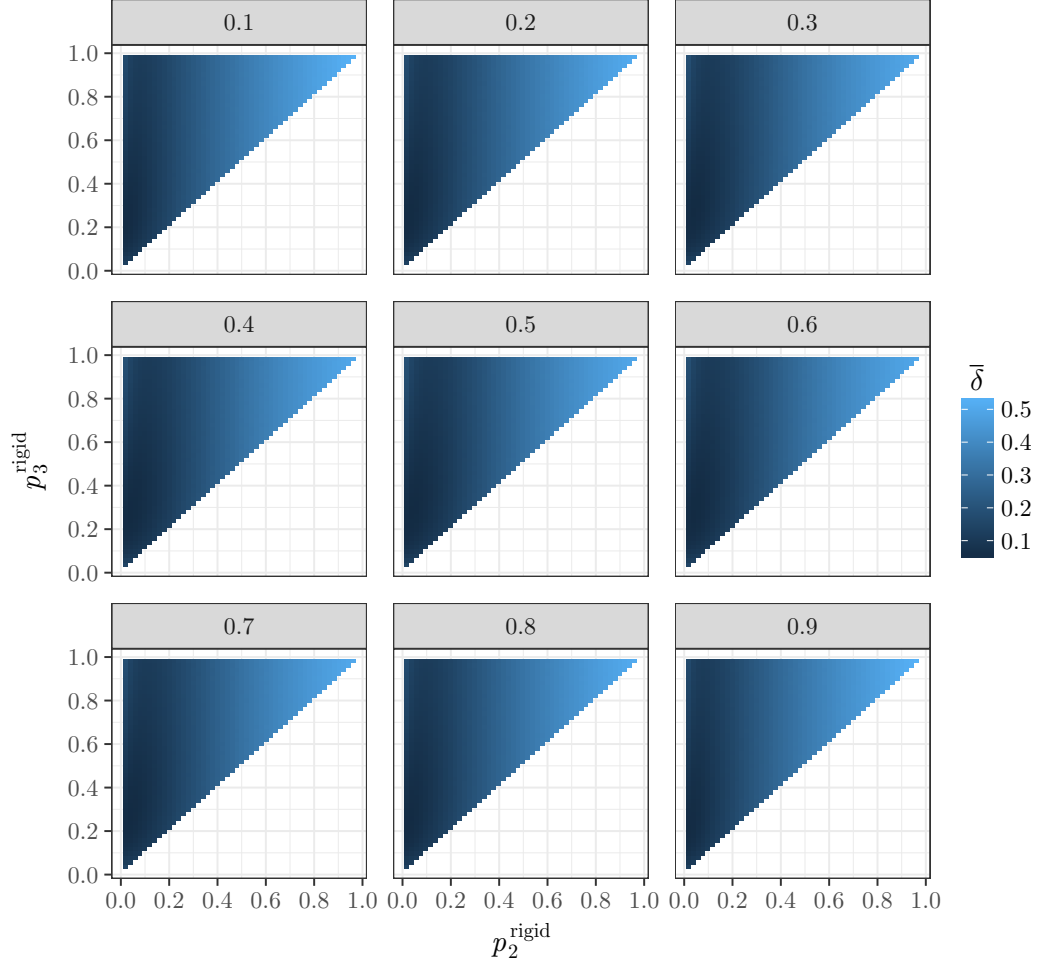


Figure 4.4. Average total variation distances for Bernoulli(p) data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, p_3^{\text{rigid}}, 1\}$. Plot labels communicate Bernoulli probabilities used.

ures 4.7 through 4.9. The average total variation distances between $\pi_{\text{target}}(\theta|\mathbf{x})$ and $\pi_{\text{rigid}}(\theta|\mathbf{x})$ have reduced from those where $r = 3$ and are now suitably small, on the order of a 5% worst-case error, and the high percentiles are also small enough for each model that we chose the corresponding odds $\mathbf{O}^{\text{rigid}} = \{1, 3, 25, 10^6\}$ for the app.

4.3.2.7 Rigid Metropolis results for $r > 4$ rigid transition probabilities. As r increases, the average total variation distance between the target posterior and that found using the rigid Metropolis sampler decreases. In fact, as $r \rightarrow \infty$, the rigid Metropolis sampler converges to the standard Metropolis procedure, which is well-

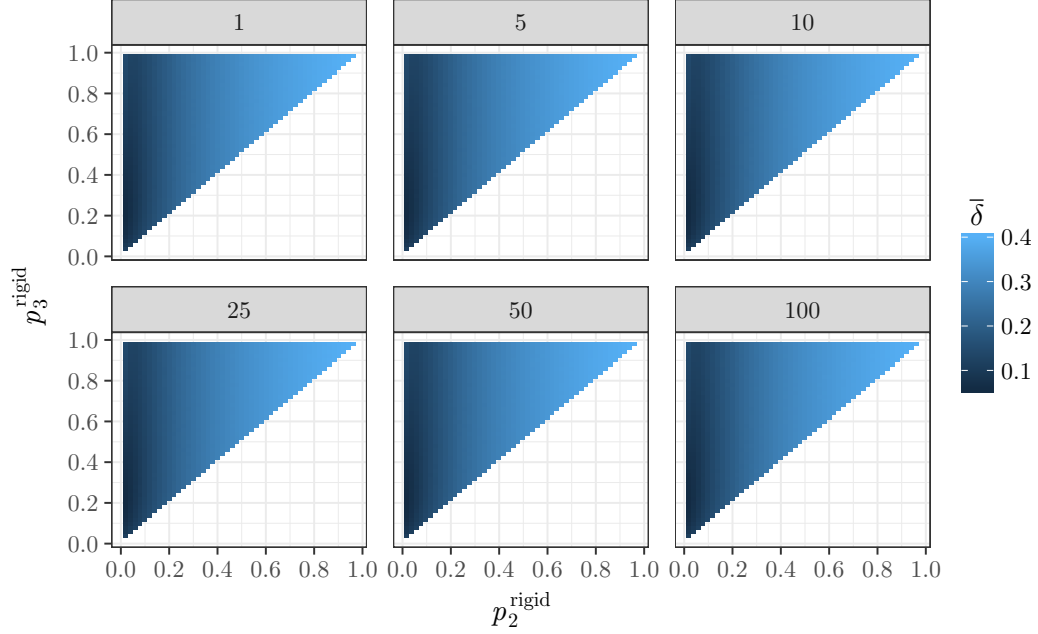


Figure 4.5. Average total variation distances for $\text{Poisson}(\lambda)$ data models based on rigid acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, p_3^{\text{rigid}}, 1\}$. Plot labels communicate Poisson rates used.

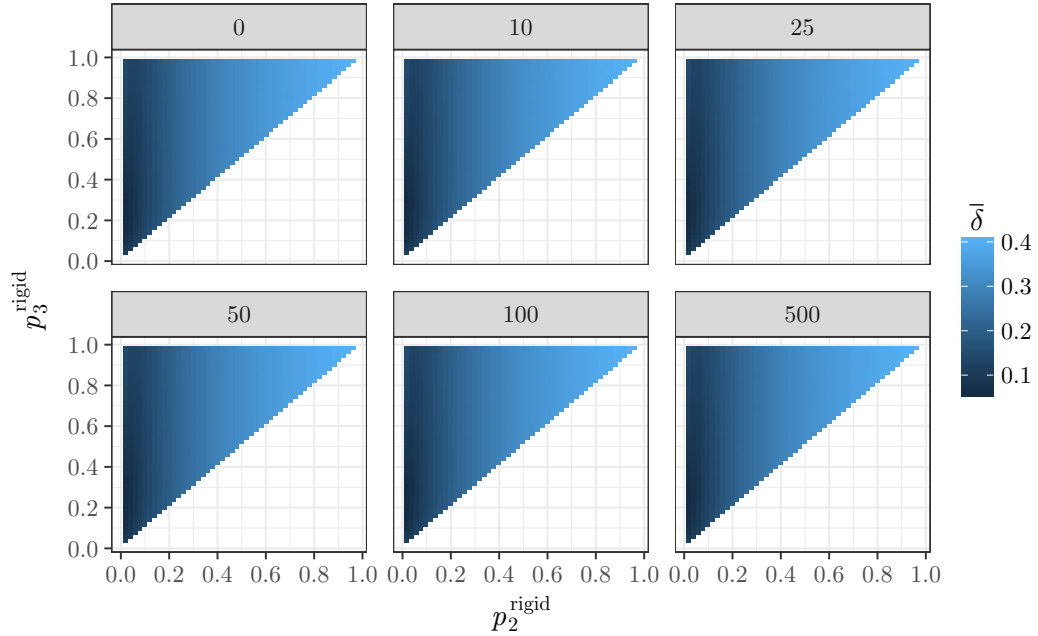


Figure 4.6. Average total variation distance for $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, p_2^{\text{rigid}}, p_3^{\text{rigid}}, 1\}$. Plot labels communicate Normal means used.

Table 4.3. Total variation distance summaries for Bernoulli(p), Poisson(λ), and $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$, indicating that four rigid probabilities improve on three, leaving a worst-case error on the order of 5%.

Data Model	Parameter	Mean δ	SD of δ	95% CI for δ
Bernoulli(p)	0.1	0.0546	0.0122	(0.0314, 0.0788)
	0.2	0.0518	0.0121	(0.0275, 0.0746)
	0.3	0.0508	0.0131	(0.0263, 0.0791)
	0.4	0.0500	0.0126	(0.0265, 0.0750)
	0.5	0.0496	0.0130	(0.0264, 0.0768)
	0.6	0.0497	0.0129	(0.0243, 0.0748)
	0.7	0.0508	0.0127	(0.0265, 0.0760)
	0.8	0.0524	0.0123	(0.0296, 0.0766)
	0.9	0.0536	0.0123	(0.0313, 0.0779)
Poisson(λ)	1	0.0747	0.0181	(0.0402, 0.1092)
	5	0.0744	0.0183	(0.0388, 0.1100)
	10	0.0749	0.0188	(0.0384, 0.1098)
	25	0.0757	0.0184	(0.0376, 0.1086)
	50	0.0740	0.0185	(0.0377, 0.1106)
	100	0.0753	0.0189	(0.0388, 0.1119)
$\mathcal{N}(\mu, 100)$	0	0.0733	0.0154	(0.0418, 0.1036)
	10	0.0740	0.0151	(0.0450, 0.1039)
	25	0.0746	0.0156	(0.0447, 0.1059)
	50	0.0736	0.0146	(0.0443, 0.1020)
	100	0.0727	0.0149	(0.0431, 0.1019)
	500	0.0734	0.0143	(0.0453, 0.1006)

known to produce chains that converge to the target posteriors. In terms of the total variation distance, this means that

$$\begin{aligned}
\delta(\pi_{\text{target}}(\theta|\mathbf{x}), \pi_{\text{MH}}(\theta|\mathbf{x})) &= \sup_{\theta \in \Theta} |\pi_{\text{target}}(\theta|\mathbf{x}) - \pi_{\text{MH}}(\theta|\mathbf{x})| \\
&= \frac{1}{2} \int_{\Theta} |f_{\text{target}}(\theta|\mathbf{x}) - f_{\text{MH}}(\theta|\mathbf{x})| d\theta \\
&\rightarrow 0
\end{aligned}$$

for posteriors $\pi_{\text{target}}(\theta|\mathbf{x})$ and $\pi_{\text{MH}}(\theta|\mathbf{x})$, the corresponding probability density functions $f_{\text{target}}(\theta|\mathbf{x})$ and $f_{\text{MH}}(\theta|\mathbf{x})$, and parameter $\theta \in \Theta$.

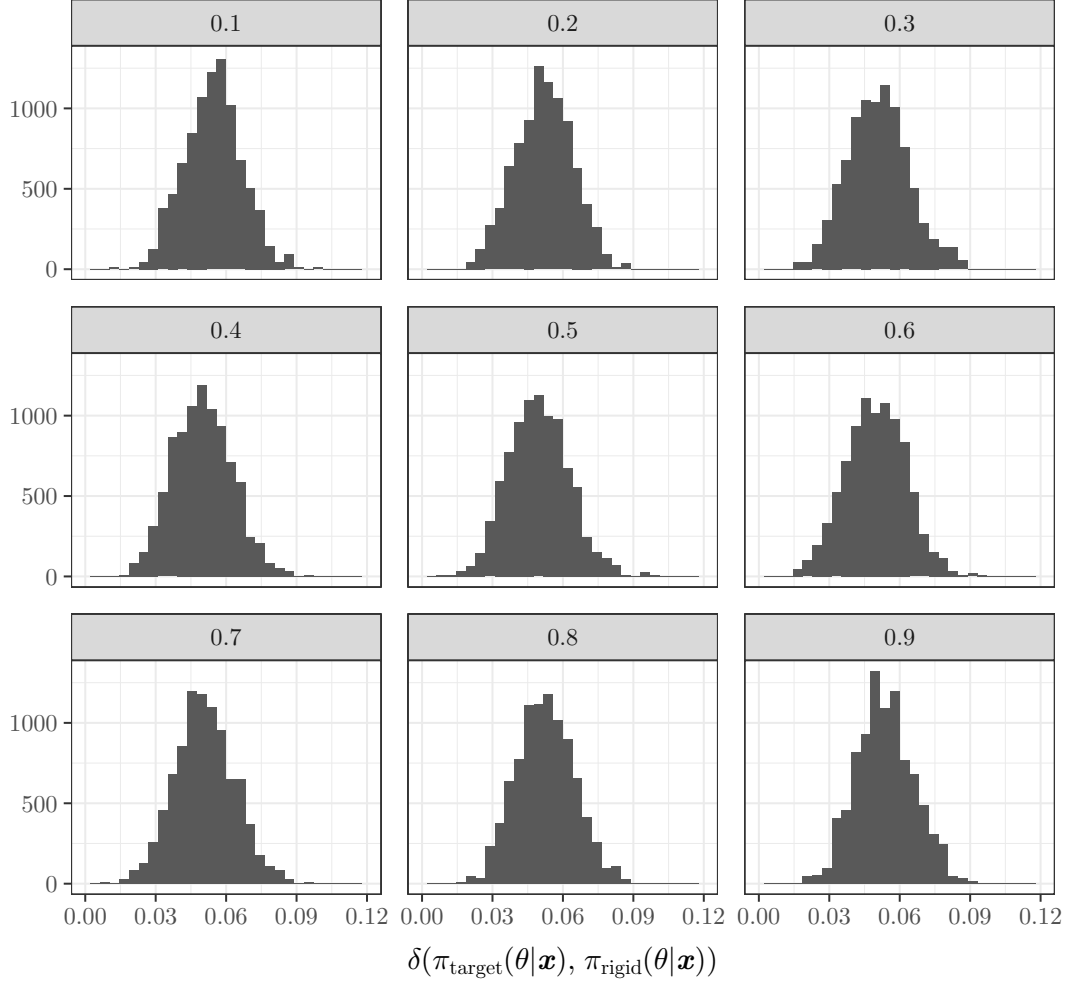


Figure 4.7. Histograms of 1,000 total variation distances for Bernoulli(p) data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$. Plot labels communicate Bernoulli probabilities used.

However, as r increases, so does the number of odds the expert must choose from in the Shiny app. As a result, it becomes increasingly difficult for them to accurately distinguish among the possible choices and, as we have seen, unnecessarily increases the complexity of the interactive process, undermining the entire process. Having obtained acceptably low average total variation distances for $r = 4$, we did not pursue probability vectors of length $r > 4$. However, the same strategy can be implemented in that area. We note in passing that while we approximated the optimal solutions using a grid-search strategy, nothing precludes the use of other, more

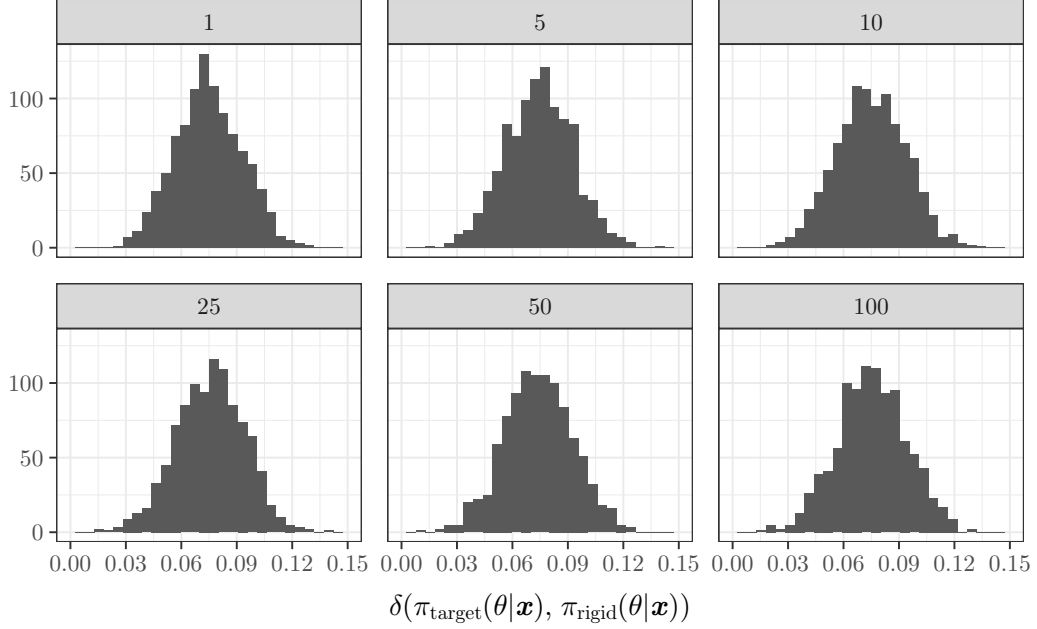


Figure 4.8. Histograms of 1,000 total variation distances for $\text{Poisson}(\lambda)$ data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$. Plot labels communicate Poisson rates used.

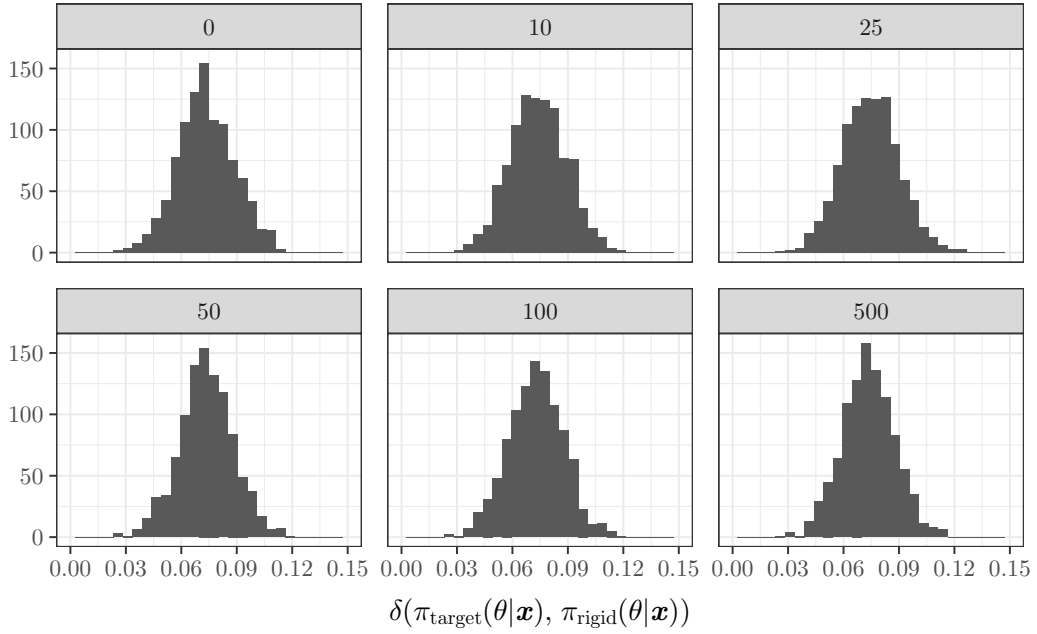


Figure 4.9. Histograms of 1,000 total variation distances for $\mathcal{N}(\mu, \sigma^2 = 100)$ data models based on acceptance probabilities $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$. Plot labels communicate Normal means used.

efficient optimization strategies. We used grid search because we were interested not only in the optimal TVD value, but also how the TVD varies across the space of rigid probabilities.

4.4 Conclusion

The graphical methods presented in Chapter 3 are based on rigid Metropolis, a variation on standard Metropolis where only a finite number of transition probabilities are used. To assess the accuracy of the rigid MCMC process, we used the total variation metric to measure the distance between the target posterior distribution and that found using a rigid MCMC. We found $\mathcal{P} = \{10^{-6}, 0.04, 0.34, 1\}$ struck a suitable balance for both the rigid procedure and the Shiny app in achieving the following goals: (1) keeping r at a small value, (2) selecting odds experts can reliably attest to, (3) maintaining the theoretical foundation of Metropolis-Hastings, and (4) attaining a sufficiently small average total variation distance.

CHAPTER FIVE

Conclusion

Prior specification is a fundamental component of any Bayesian analysis. In situations where expert opinion is desired, prior elicitation becomes a necessary task. While various methods and tools exist that enable experts to incorporate their belief about a parameter in a statistical analysis, the methods require the expert to quantify summaries they may not be able to reliably attest to. Additionally, important steps such as training and post-elicitation adequacy checks for the prior are often undervalued if not overlooked entirely.

In this dissertation we proposed graphical elicitation procedures that address the drawbacks to existing methods by enabling the expert to work on an observational scale rather than directly with parameters. We also focus on the elicitation process in a holistic way. We train the expert in the natural variability in datasets by leading them through a Rorschach test. Once they have become more comfortable visualizing and understanding such variability for the data model at hand, the expert proceeds to the elicitation stage, where they make a series of lineup-style selections of graphics of hypothetical future datasets. A prior is then fit to these selections through a conversion process, and the expert is asked to assess the resulting prior by examining a density plot and summaries.

The first procedure, which is deterministic in nature, relies on the expert's selections in addition to the effective prior sample size they specify. The second procedure, on the other hand, is built on a stochastic transitioning scheme and fits a prior directly to the parameter values associated with the expert's selections. The transitioning mechanism is based on simulation results from rigid MCMCs, variations

on the standard Metropolis algorithm that permit discrete sets of acceptance probabilities.

The methods proposed are novel schemes that have opened doors to new lines of research for prior elicitation. Future areas of research include applications to other univariate data models (such as the Student t), multivariate data models, and regression models, as well as new graphical elicitation schemes built on other transitioning structures (e.g. rejection sampling and variations on those proposed in this dissertation).

APPENDICES

APPENDIX A

Effective Sample Size Computations

A.1 Data Model and Prior Combinations with Closed Form ESS Expressions

Closed form expressions for the effective sample size (ESS) of a prior exist for various data model and prior combinations, among them three of the four discussed in this dissertation: the Bernoulli(p), Poisson(λ), and $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known, all with conjugate priors. Expressions for the three cases can be found in Table A.1.

Table A.1. Effective sample size expressions for three common cases.

Data Model	Prior	ESS
Bernoulli(p)	Beta(α, β)	$\alpha + \beta$
Poisson(λ)	Gamma(α, β)	β
$\mathcal{N}(\mu, \sigma^2), \sigma^2$ known	$\mathcal{N}(\mu_0, \sigma_0^2)$	σ^2 / σ_0^2

A.2 $\mathcal{N}(\mu, \sigma^2)$ Data Model with μ and σ^2 Unknown and a Normal-Inverse-Gamma Joint Prior

While closed form expressions for the ESS of a prior exist for various data models and prior combinations, such as the three included in Appendix A.1, the ESS does not have a closed form expression when using various combinations of data models and priors and must instead be approximated via simulation. For example, such simulations are necessary for computing the ESS of the joint Normal-Inverse-Gamma($\tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}$) prior on $[\mu, \sigma^2]$ when assumed with the $\mathcal{N}(\mu, \sigma^2)$ data model with μ and σ^2 unknown. We now define the expressions required to run the simulations as presented in Morita et al. (2008) and then apply them to this data model and prior combination:

- $p(\boldsymbol{\theta} | \tilde{\boldsymbol{\theta}}) = \text{prior on } \boldsymbol{\theta}$

- $q_0(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0) = \epsilon$ -information prior on $\boldsymbol{\theta}$: a prior with a variance inflated by a large constant c
- $q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m) =$ posterior on $\boldsymbol{\theta}$ given a dataset \mathbf{Y} of size m
- $D_{p,j}(\boldsymbol{\theta}) = -\frac{\partial^2 \log p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})}{\partial \theta_j^2}, \quad j = 1, \dots, d$
- $D_{q,j}(m, \boldsymbol{\theta}, \mathbf{Y}_m) = -\frac{\partial^2 \log q_m(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}_0, \mathbf{Y}_m)}{\partial \theta_j^2}$
- $\bar{\boldsymbol{\theta}} = E_p(\boldsymbol{\theta}) =$ prior mean under $p(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$
- $D_{p,+}(\boldsymbol{\theta}) = \sum_{j=1}^d D_{p,j}(\boldsymbol{\theta})$
- $D_{q,+}(m, \boldsymbol{\theta}) = \sum_{j=1}^d \int D_{q,j}(m, \boldsymbol{\theta}, \mathbf{Y}_m) f_m(\mathbf{Y}_m) d\mathbf{Y}_m$
- $\delta(m, \bar{\boldsymbol{\theta}}, p, q_0) = |D_{p,+}(\bar{\boldsymbol{\theta}}) - D_{q,+}(m, \bar{\boldsymbol{\theta}})|$

In order to compute the ESS of the joint prior, we must find expressions for $D_{p,+}(\bar{\boldsymbol{\theta}})$ and $D_{q,+}(m, \bar{\boldsymbol{\theta}})$.

We start by finding $D_{p,+}(\bar{\boldsymbol{\theta}})$. For a Normal-inverse-gamma $(\tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta})$ prior on $\boldsymbol{\theta} = [\mu, \sigma^2]$, which can be decomposed into the hierarchical prior structure

$$\begin{aligned} \mu|\sigma^2, \tilde{\mu}, \tilde{\lambda} &\sim \mathcal{N}\left(\tilde{\mu}, \frac{\sigma^2}{\tilde{\lambda}}\right) \\ \sigma^2|\tilde{\alpha}, \tilde{\beta} &\sim \text{Inverse-Gamma}(\tilde{\alpha}, \tilde{\beta}), \end{aligned}$$

the probability density function is

$$p\left(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}\right) = \frac{\sqrt{\tilde{\lambda}}}{\sqrt{2\pi\sigma^2}} \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \left(\frac{1}{\sigma^2}\right)^{\tilde{\alpha}+1} \exp\left\{-\frac{2\tilde{\beta} + \tilde{\lambda}(\mu - \tilde{\mu})^2}{2\sigma^2}\right\}. \quad (\text{A.1})$$

(A.1) leads to

$$\begin{aligned} \log p\left(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}\right) &= \frac{1}{2} \left(\log \tilde{\lambda} - \log(2\pi) - \log \sigma^2 \right) + \tilde{\alpha} \log \tilde{\beta} - \log \Gamma(\tilde{\alpha}) \\ &\quad - (\tilde{\alpha} + 1) \log \sigma^2 - \frac{1}{2\sigma^2} \left(2\tilde{\beta} + \tilde{\lambda}(\mu - \tilde{\mu})^2 \right). \end{aligned}$$

Thus,

$$\frac{\partial \log p \left(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta} \right)}{\partial \mu} = -\frac{\tilde{\lambda}(\mu - \tilde{\mu})}{\sigma^2},$$

which implies

$$\frac{\partial^2 \log p \left(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta} \right)}{\partial \mu^2} = -\frac{\tilde{\lambda}}{\sigma^2}.$$

We then have

$$D_{p,1}(\mu, \sigma^2) = -\frac{\partial^2 \log p \left(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta} \right)}{\partial \mu^2} = \frac{\tilde{\lambda}}{\sigma^2}. \quad (\text{A.2})$$

Next, (A.1) leads to

$$\frac{\partial \log p \left(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta} \right)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} - \frac{\tilde{\alpha} + 1}{\sigma^2} + \frac{2\tilde{\beta} + \tilde{\lambda}(\mu - \tilde{\mu})^2}{2(\sigma^2)^2},$$

which implies

$$\frac{\partial^2 \log p \left(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta} \right)}{\partial (\sigma^2)^2} = \frac{1}{2(\sigma^2)^2} + \frac{\tilde{\alpha} + 1}{(\sigma^2)^2} - \frac{2\tilde{\beta} + \tilde{\lambda}(\mu - \tilde{\mu})^2}{(\sigma^2)^3}.$$

Thus

$$\begin{aligned} D_{p,2}(\mu, \sigma^2) &= -\frac{\partial^2 \log p \left(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta} \right)}{\partial (\sigma^2)^2} \\ &= -\frac{\tilde{\alpha} + \frac{3}{2}}{(\sigma^2)^2} + \frac{2\tilde{\beta} + \tilde{\lambda}(\mu - \tilde{\mu})^2}{(\sigma^2)^3}. \end{aligned} \quad (\text{A.3})$$

(A.2) and (A.3) result in

$$\begin{aligned} D_{p,+}(\mu, \sigma^2) &= \sum_{j=1}^2 D_{p,j}(\mu, \sigma^2) \\ &= \frac{\tilde{\lambda}}{\sigma^2} - \frac{\tilde{\alpha} + \frac{3}{2}}{(\sigma^2)^2} + \frac{2\tilde{\beta} + \tilde{\lambda}(\mu - \tilde{\mu})^2}{(\sigma^2)^3}. \end{aligned}$$

Evaluating $D_{p,+}(\mu, \sigma^2)$ at the prior mean,

$$\bar{\theta} = E_p[\mu, \sigma^2] = \left[\tilde{\mu}, \frac{\tilde{\beta}}{\tilde{\alpha} - 1} \right], \quad (\text{A.4})$$

we have

$$\begin{aligned} D_{p,+}(\bar{\boldsymbol{\theta}}) &= D_{p,+}\left(\tilde{\mu}, \frac{\tilde{\beta}}{\tilde{\alpha}-1}\right) \\ &= \frac{\tilde{\lambda}}{\sigma_*^2} - \frac{\tilde{\alpha} + \frac{3}{2}}{(\sigma_*^2)^2} + \frac{2\tilde{\beta}}{(\sigma_*^2)^3} \end{aligned} \quad (\text{A.5})$$

where

$$\sigma_*^2 = \frac{\tilde{\beta}}{\tilde{\alpha}-1}.$$

The next step is to find $D_{q,+}(m, \bar{\boldsymbol{\theta}})$, which first requires determining the posterior, $q_m(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \mathbf{y}_m)$, on $[\mu, \sigma^2]$ for a sample \mathbf{y} of size m and with ϵ -information prior $q_0(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}_0, \tilde{\beta}_0, c)$.

For the Normal-inverse-gamma $(\tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta})$ prior on $\boldsymbol{\theta} = [\mu, \sigma^2]$, the associated ϵ -information prior can be written hierarchically as

$$\begin{aligned} \mu \mid \sigma^2, \tilde{\mu}, \tilde{\lambda}, c &\sim \mathcal{N}\left(\tilde{\mu}, \frac{c\sigma^2}{\tilde{\lambda}}\right) \\ \sigma^2 \mid \tilde{\alpha}_0, \tilde{\beta}_0, c &\sim \text{Inverse-Gamma}\left(\tilde{\alpha}_0, \tilde{\beta}_0\right), \end{aligned}$$

where

$$\tilde{\alpha}_0 = 2 + (2c)^{-1} \quad \text{and} \quad \tilde{\beta}_0 = \frac{(4 + c^{-1})\tilde{\nu}\tilde{\sigma}^2}{4(\tilde{\nu} - 2)}.$$

We then obtain the posterior pdf:

$$\begin{aligned} q_m(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \mathbf{y}_m) &= \left[\prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\} \right] \\ &\times \frac{\sqrt{\tilde{\lambda}}}{\sqrt{2\pi c\sigma^2}} \exp\left\{-\frac{\tilde{\lambda}(\mu - \tilde{\mu})^2}{2c\sigma^2}\right\} \\ &\times \frac{\tilde{\beta}_0^{\tilde{\alpha}_0}}{\Gamma(\tilde{\alpha}_0)} \left(\frac{1}{\sigma^2}\right)^{\tilde{\alpha}_0+1} \exp\left\{-\frac{\tilde{\beta}_0}{\sigma^2}\right\}. \end{aligned}$$

Next,

$$\begin{aligned} q_m(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \mathbf{y}_m) &= (2\pi)^{-\frac{1}{2}(m+1)} \left(\frac{\tilde{\lambda}}{c}\right)^{\frac{1}{2}} \frac{\tilde{\beta}_0^{\tilde{\alpha}_0}}{\Gamma(\tilde{\alpha}_0)} (\sigma^2)^{-\frac{1}{2}(m+2\tilde{\alpha}_0+3)} \\ &\times \exp\left\{-\frac{\sum_{i=1}^m (y_i - \mu)^2 + \frac{\tilde{\lambda}}{c}(\mu - \tilde{\mu})^2 + 2\tilde{\beta}_0}{2\sigma^2}\right\}, \end{aligned}$$

which implies

$$\begin{aligned} \log q_m(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \mathbf{y}_m) &= -\frac{1}{2}(m+1)\log(2\pi) + \frac{1}{2}(\log \tilde{\lambda} - \log c) \\ &\quad + \tilde{\alpha}_0 \log \tilde{\beta}_0 - \log \Gamma(\tilde{\alpha}_0) - \frac{1}{2}(m+2\tilde{\alpha}_0+3)\log \sigma^2 \\ &\quad - \frac{1}{2\sigma^2} \left(\sum_{i=1}^m (y_i - \mu)^2 + \frac{\tilde{\lambda}}{c}(\mu - \tilde{\mu})^2 + 2\tilde{\beta}_0 \right). \end{aligned} \quad (\text{A.6})$$

Thus,

$$\frac{\partial \log q_m(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \mathbf{y}_m)}{\partial \mu} = \frac{1}{\sigma^2} \left(\sum_{i=1}^m (y_i - \mu) - \frac{\tilde{\lambda}}{c}(\mu - \tilde{\mu}) \right),$$

which leads to

$$\frac{\partial^2 \log q_m(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta}, \mathbf{y}_m)}{\partial \mu^2} = -\frac{mc + \tilde{\lambda}}{c\sigma^2}.$$

We then have

$$\begin{aligned} D_{q,1}(m, \mu, \sigma^2, \mathbf{y}_m) &= -\frac{\partial^2 \log q_m(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta})}{\partial \mu^2} \\ &= \frac{mc + \tilde{\lambda}}{c\sigma^2}. \end{aligned} \quad (\text{A.7})$$

Next, (A.6) implies

$$\frac{\partial \log q_m}{\partial \sigma^2} = -\frac{m+2\tilde{\alpha}_0+3}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \left(\sum_{i=1}^m (y_i - \mu)^2 + \frac{\tilde{\lambda}}{c}(\mu - \tilde{\mu})^2 + 2\tilde{\beta}_0 \right),$$

which leads to

$$\frac{\partial^2 \log q_m}{\partial (\sigma^2)^2} = \frac{m+2\tilde{\alpha}_0+3}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \left(\sum_{i=1}^m (y_i - \mu)^2 + \frac{\tilde{\lambda}}{c}(\mu - \tilde{\mu})^2 + 2\tilde{\beta}_0 \right).$$

Thus

$$\begin{aligned} D_{q,2}(m, \mu, \sigma^2, \mathbf{y}_m) &= -\frac{\partial^2 \log q_m(\mu, \sigma^2 \mid \tilde{\mu}, \tilde{\lambda}, \tilde{\alpha}, \tilde{\beta})}{\partial (\sigma^2)^2} \\ &= -\frac{m+2\tilde{\alpha}_0+3}{2(\sigma^2)^2} \\ &\quad + \frac{1}{(\sigma^2)^3} \left(\sum_{i=1}^m (y_i - \mu)^2 + \frac{\tilde{\lambda}}{c}(\mu - \tilde{\mu})^2 + 2\tilde{\beta}_0 \right). \end{aligned} \quad (\text{A.8})$$

(A.7) and (A.8) next result in

$$\begin{aligned}
D_{q,+}(m, \boldsymbol{\theta}) &= \sum_{j=1}^2 \int D_{q,j}(m, \boldsymbol{\theta}, \mathbf{y}_m) f_m(\mathbf{y}_m) d\mathbf{y}_m \\
&= \int \frac{mc + \tilde{\lambda}}{c\sigma^2} f_m(\mathbf{y}_m) d\mathbf{y}_m - \int \frac{m + 2\tilde{\alpha}_0 + 3}{2(\sigma^2)^2} f_m(\mathbf{y}_m) d\mathbf{y}_m \\
&\quad + \int \frac{1}{(\sigma^2)^3} \left(\sum_{i=1}^m (y_i - \mu)^2 + \frac{\tilde{\lambda}}{c} (\mu - \tilde{\mu})^2 + 2\tilde{\beta}_0 \right) f_m(\mathbf{y}_m) d\mathbf{y}_m \\
&= \frac{mc + \tilde{\lambda}}{c\sigma^2} - \frac{m + 2\tilde{\alpha}_0 + 3}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \left(\frac{\tilde{\lambda}}{c} (\mu - \tilde{\mu})^2 + 2\tilde{\beta}_0 \right) \\
&\quad + \frac{1}{(\sigma^2)^3} \int \sum_{i=1}^m (y_i - \mu)^2 f_m(\mathbf{y}_m) d\mathbf{y}_m
\end{aligned}$$

Now recall c is a large constant. This implies that

$$\begin{aligned}
D_{q,+}(m, \boldsymbol{\theta}) &\approx \frac{m}{\sigma^2} - \frac{m + 2\tilde{\alpha}_0 + 3}{2(\sigma^2)^2} \\
&\quad + \frac{1}{(\sigma^2)^3} \left(2\tilde{\beta}_0 + \int \sum_{i=1}^m (y_i - \mu)^2 f_m(\mathbf{y}_m) d\mathbf{y}_m \right).
\end{aligned}$$

Evaluating $D_{q,+}(m, \boldsymbol{\theta})$ at the prior mean specified previously in (A.4) leads to

$$\begin{aligned}
D_{q,+}(m, \bar{\boldsymbol{\theta}}) &= \frac{m}{\sigma_\star^2} - \frac{m + 2\tilde{\alpha}_0 + 3}{2(\sigma_\star^2)^2} \\
&\quad + \frac{1}{(\sigma_\star^2)^3} \left(2\tilde{\beta}_0 + \int \sum_{i=1}^m (y_i - \tilde{\mu})^2 f_m(\mathbf{y}_m) d\mathbf{y}_m \right),
\end{aligned} \tag{A.9}$$

where

$$\sigma_\star^2 = \frac{\tilde{\beta}}{\tilde{\alpha} - 1}.$$

Next, from (A.5) and (A.9) we have

$$\begin{aligned}
\delta(m, \bar{\boldsymbol{\theta}}, p, q_0) &= \left| D_{p,+}(\bar{\boldsymbol{\theta}}) - D_{q,+}(m, \bar{\boldsymbol{\theta}}) \right| \\
&= \left| \frac{\tilde{\lambda} - m}{\sigma_\star^2} - \frac{\tilde{\alpha} - \frac{m}{2} - \tilde{\alpha}_0}{(\sigma_\star^2)^2} + \frac{2(\tilde{\beta} - \tilde{\beta}_0)}{(\sigma_\star^2)^3} \right. \\
&\quad \left. - \frac{1}{(\sigma_\star^2)^3} \int \sum_{i=1}^m (y_i - \tilde{\mu})^2 f_m(\mathbf{y}_m) d\mathbf{y}_m \right|.
\end{aligned}$$

We finally obtain the effective sample size:

$$\text{ESS} = \underset{m}{\text{argmin}} \delta(m, \bar{\boldsymbol{\theta}}, p, q_0).$$

BIBLIOGRAPHY

- Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., and Hyndman, R. (2016). *rmarkdown: Dynamic Documents for R*. R package version 1.3.
- Blair, S. (2017). *Contributions to the theory and practice of prior elicitation in biopharmaceutical research*. PhD thesis, Baylor University.
- Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383.
- Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2016). *shiny: Web Application Framework for R*. R package version 0.13.2.
- Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011). *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC Press.
- Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D., and Ushey, K. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, 40(8):1–18.
- Garthwaite, P. H., Kadane, J. B., and O’Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*. Chapman & Hall/CRC Boca Raton, FL, USA, 3 edition.
- Jackman, S. (2009). *Bayesian analysis for the social sciences*, volume 846. John Wiley & Sons.
- Jones, G. and Johnson, W. O. (2014). Prior elicitation: interactive spreadsheet graphics with sliders can be fun, and informative. *The American Statistician*, 68(1):42–51.
- Kahle, D. and Stamey, J. (2016). *invgamma: The Inverse Gamma Distribution*. R package version 1.0.
- Kahle, D., Stamey, J., Natanegara, F., Price, K., and Han, B. (2014). Facilitated prior elicitation with the wolfram cdf. *JSM Proceedings*.
- Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77(3):217–273.

- Lichtenstein, S. and Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational behavior and human performance*, 20(2):159–183.
- Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In *Decision making and change in human affairs*, pages 275–324. Springer.
- Morita, S., Thall, P. F., and Müller, P. (2008). Determining the effective sample size of a parametric prior. *Biometrics*, 64(2):595–602.
- Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52:1–4.
- Oakley, J. and O’Hagan, A. (2010). SHELF: the Sheffield Elicitation Framework (version 2.0). *Sheffield, UK: School of Mathematics and Statistics, University of Sheffield*.
- O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: eliciting experts’ probabilities*. John Wiley & Sons.
- Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of consulting psychology*, 29(3):261.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Resnick, S. I. (2013). *A probability path*. Springer Science & Business Media.
- Ruckdeschel, P., Kohl, M., Stabla, T., and Camphausen, F. (2006). S4 classes for distributions. *R News*, 6(2):2–6.
- Schaefer, R. E. and Borcharding, K. (1973). The assessment of subjective probability distributions: A training experiment. *Acta Psychologica*, 37(2):117–129.
- Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, pages 605–610.
- Su, C.-L. (2006). BetaBuster. Department of Medicine and Epidemiology, University of California, Davis.
- Van den Steen, E. (2011). Overconfidence by bayesian-rational agents. *Management Science*, 57(5):884–896.
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

- Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.
- Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010). Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):973–979.
- Wickham, H. and Francois, R. (2016). *dplyr: A Grammar of Data Manipulation*. R package version 0.5.0.
- Winkler, R. L. (1967). The assessment of prior distributions in bayesian analysis. *Journal of the American Statistical Association*, 62(319):776–800.
- Wu, Y., Shih, W. J., and Moore, D. F. (2008). Elicitation of a beta prior for bayesian inference in clinical trials. *Biometrical Journal*, 50(2):212–223.