

HARVARD UNIVERSITY  
Graduate School of Arts and Sciences



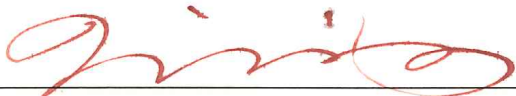
DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the  
Department of Statistics  
have examined a dissertation entitled


**Information: measuring the missing, using the observed,  
and approximating the complete**

presented by **David E. Jones**

candidate for the degree of Doctor of Philosophy and hereby  
certify that it is worthy of acceptance.

Signature  \_\_\_\_\_

Typed name: Professor Xiao-Li Meng

Signature  \_\_\_\_\_

Typed name: Professor David A. van Dyk

Signature  \_\_\_\_\_

Typed name: Professor Tirthankar Dasgupta

Date: March 29, 2016



Information: measuring the missing,  
using the observed, and approximating  
the complete

A dissertation presented  
by

David E. Jones

to

The Department of Statistics  
in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy  
in the subject of

Statistics

Harvard University  
Cambridge, Massachusetts  
March 2016

©2016 – David E. Jones

All rights reserved.

## Information: measuring the missing, using the observed, and approximating the complete

### Abstract

In this thesis, we present three topics broadly connected to the concept and use of statistical information, and specifically regarding the problems of hypothesis testing and model selection, astronomical image analysis, and Monte Carlo integration.

The first chapter is inspired by the work of DeGroot (1962) and Nicolae et al. (2008) and is the most directly focused on the theme of statistical information. DeGroot (1962) developed a general framework for constructing Bayesian measures of the expected information that an experiment will provide for estimation. We propose an analogous framework for measures of information for hypothesis testing, and illustrate how these measures can be applied in experimental design. In contrast to estimation information measures that are typically used in experimental design for surface estimation, test information measures are more useful in experimental design for hypothesis testing and model selection. Indeed, one test information measure suggested by our framework is probability based, and in design contexts where decision problem are of interest, it has more appealing properties than variance based measures. The underlying intuition of our de-

sign proposals is straightforward: to distinguish between two or more models we should collect data from regions of the covariate space for which the models differ most. [Nicolae et al. \(2008\)](#) give an asymptotic equivalence between their test information measures and Fisher information. We extend this result to all test information measures under our framework, and hence further our understanding of the links between test and estimation information measures.

In the second chapter, we present a powerful new algorithm that combines both spatial and spectral (energy) information to separate photons from overlapping sources (e.g., stars) in an astronomical image. We use Bayesian statistical methods to simultaneously infer the number of overlapping sources, to probabilistically separate the photons among the sources, and to fit the parameters describing the individual sources. Using the Bayesian joint posterior distribution, we are able to coherently quantify the uncertainties associated with all these parameters. The advantages of combining spatial and spectral information are demonstrated through a simulation study. The utility of the approach is then illustrated by analysis of observations of the sources FK Aqr and FL Aqr with the *XMM-Newton* Observatory and the central region of the Orion Nebula Cluster with the *Chandra* X-ray Observatory. In this chapter we make additional effort to explain relevant standard statistical ideas and methods in order to make the exposition more accessible to astronomers unfamiliar with statistics.

The last chapter extends the maximum likelihood theory developed by [Kong et al. \(2003\)](#) for deriving Monte Carlo estimators of normalizing constants. [Kong et al. \(2003\)](#) had the fundamental idea of treating the baseline measure as an

unknown quantity to be estimated, and found that this suggested a maximum likelihood method for estimating integrals of interest. Their work shows that sub-models of the baseline measure can be used to incorporate some of our knowledge of the true measure, thus allowing greater statistical precision to be gained at the expense of more function evaluations, but without the need for more Monte Carlo samples. Our contribution is to introduce a simple extension of this framework which greatly increases its flexibility for trading off statistical and computational efficiency. As a result, we gain an appealing maximum likelihood interpretation of the very effective warp transformations proposed by [Meng and Schilling \(2002\)](#). We additionally investigate the open problem of optimally choosing parameters for sub-models of the baseline measure.

## Citation to previously published work

### Chapter 2:

The material in Chapter 2 is published in the *Astrophysical Journal* (ApJ):

D. E. Jones, V. L. Kashyap, D. A. van Dyk. Disentangling Overlapping Astronomical Sources using Spatial and Spectral Information. *Astrophysical Journal*, 808(3):137 – 160, 2015. DOI: 10.1088/0004-637X/808/2/137.

The American Astronomical Society (AAS) holds the copyright for all material published in its journals (including ApJ) but permits authors to reproduce their own work within their lifetimes, as stated in the AAS copyright document: <http://iopscience.iop.org/0004-637X/page/Copyright+and+permissions>.

### Chapters 1 and 3:

The material in Chapters 1 and 3 has not yet been published elsewhere.



# Contents

<b>1</b>	<b>Designing test information and test information in design</b>	<b>1</b>
1.1	Motivation and overview . . . . .	1
1.2	Expected test information: general framework and applications . . . . .	13
1.3	Observed test information in theory and application . . . . .	31
1.4	Links between test and estimation information . . . . .	39
1.5	Discussion and further work . . . . .	44
<b>2</b>	<b>Disentangling overlapping astronomical sources using spatial and spectral information</b>	<b>48</b>
2.1	Introduction . . . . .	48
2.2	Data and statistical models . . . . .	55
2.3	Illustrative example . . . . .	65
2.4	Bayesian model fitting . . . . .	69
2.5	Simulation studies . . . . .	79
2.6	Application I: XMM dataset . . . . .	91
2.7	Application II: Chandra dataset . . . . .	96
2.8	Summary . . . . .	106
<b>3</b>	<b>Likelihood methods for Monte Carlo estimation</b>	<b>108</b>
3.1	Introduction . . . . .	108
3.2	Likelihood methods for optimizing Monte Carlo integration . . . . .	113
3.3	Warp transformations and beyond . . . . .	127
3.4	Choosing invariance group transformations . . . . .	136
3.5	Summary and future work . . . . .	148
<b>A</b>	<b>Appendices</b>	<b>151</b>
A.1	Proof of Theorem 1.1 . . . . .	151
A.2	Split and combine proposals in reversible jump MCMC . . . . .	152
A.3	Label switching . . . . .	159
A.4	King profile . . . . .	162
A.5	Equivalence of the ML and warp bridge sampling estimators . . . . .	163
	<b>References</b>	<b>165</b>

# List of figures

- 1.1 Summary of estimation and test information theory. The synthesis of test information measures into one coherent framework paralleling the estimation framework is new. Also new are the general links between estimation and test information, although Nicolae et al. (2008) considered the same connection with Fisher information for specific cases. . . . . 17
- 1.2 Prior mean power of the likelihood ratio test under  $M_C$ , for  $C \in \{\text{Spread, Power, } P_{\text{bayes}}, P_{\text{LRT}}, KLT_{\text{bayes}}, KLT_{\text{LRT}}\}$ , for different settings of the priors in (1.26). For the P-optimal designs we set  $\pi_0 = \pi_1 = 0.5$ . . . . . 25
- 1.3 Comparison of  $M_{\text{spread}}$  (left) and  $M_{\text{power}}$  (right), with the parameters of (1.26) set to  $\eta = \{-2, 10\}$  and  $R = 10I_2$ . Thin grey lines show  $d_t$ , for each simulated dataset, and the thick red line shows  $s_t$  (both are described in the main text). The large dots show the design point locations (x-coordinates) and the corresponding values of  $s_t$  (y-coordinates and numbers below). . . . . 27
- 1.4 The null and true cubic regression models and the observed data posterior mean fit. The observed data are indicated by large dots. The left and right plots show example simulations used in producing parts (a) and (b) of Figure 1.5, respectively. . . . . 37
- 1.5 Prior mean power of the likelihood ratio test under the conditional D-optimality, P-optimality, and KLT-optimality procedures, across 250 datasets simulated under  $\beta \sim N(\eta, 0.2I_4)$  (first row, unconstrained values). The main text describes the generation of  $\beta_0$  and  $\eta$  for parts (a) and (b). The first row constrained values show the prior mean powers when the only missing data designs allowed are (i) and (ii) (see the main text). For this case, the second row shows the percentage of simulations in which design (i) was selected. . . 38

2.1	Fitting <i>gamma</i> distributions to a counts spectrum. The histogram shows the observed spectrum of the brightest of the Chandra sources in the Orion field in Section 2.7.2 (from one iteration of our algorithm; see Section 2.4.3), and the curves show <i>gamma</i> model fits. The solid line (green) is the extended full model fit of the two- <i>gamma</i> spectral model and the dashed line (red) is the maximum likelihood fit of the one- <i>gamma</i> model. . . . .	62
2.2	Illustrative simulation setup. Locations of three weak sources are shown as red dots over a scatter plot (left), as also are the adopted counts spectra of the sources and the background (right). . . . .	66
2.3	Probability distribution of the number of sources based on the spatial-only model (left) and the full model (right). In this simulation, the true value is $K = 3$ . . . . .	67
2.4	Simulated dataset for the 10 source case. The simulated spatial counts distribution (left) and the adopted spectra for each source and the background (right) are shown. The true locations of the 10 sources are marked by large (red) dots in the left plot. . . . .	80
2.5	Average posterior probabilities of plausible values of $K$ across ten datasets. Left plots show posteriors for the ten-source reality ( $K_{\text{true}} = 10$ ) with prior mean values of $\kappa = 1, 3, 10$ from top to bottom. Right plots show posteriors for the one-source reality ( $K_{\text{true}} = 1$ ) with $\kappa = 1, 3, 10$ . In each plot, the 25% and 75% quantiles across the 10 datasets are indicated by the vertical error bars for each value of $K$ . . . . .	82
2.6	Exploring the sensitivity of our algorithm to source separation, relative strengths, and background level. The median posterior probability of $K = 2$ across the 100 simulations is shown; $K_{\text{true}} = 2$ in all cases. The results from the spatial-only model (left column) and the full model (right column) are both shown. Red indicates probabilities less than 0.1, and white indicates probabilities greater than 0.5. (Intermediate colors indicate probabilities between 0.1 and 0.5.) . . . . .	86

2.7	Sensitivity of location determination as a function of source separation, relative strength, and background level. The simulation is the same as that in Figure 2.6. Mean posterior locations of two sources for each of 100 simulations, under the spatial-only model (top 20 plots) and the full model (bottom 20 plots). Red and blue dots give the mean posterior locations for each simulation of the bright and faint sources respectively. The large ‘X’s of corresponding color indicate the true locations. The diameters of the dots are proportional to the posterior probabilities of two sources. The relative background, relative source intensity, and source separation are indicated by $b$ , $r$ and $d$ respectively. . . . .	89
2.8	Visual binary FK and FL Aqr observed with XMM-Newton (FK is the brighter source at bottom). The XMM obs_id is 0151450101. Shown is a counts image with $10''$ bins and arbitrary origin (left), and a scatter plot of a subset of 6,000 events over a 5ks subexposure (right). . . . .	93
2.9	A histogram of the spectral data in the XMM observation of FK Aqr and FL Aqr. Plotted are 1,000 spectra for the bright (solid black lines) and faint (dashed red lines) sources, each corresponds to a posterior sample of the spectral parameters. (The posterior variance is small on this scale.) The background spectra is shown by the dotted green line. . . . .	93
2.10	Posterior distributions of the parameters of the <i>gamma</i> distributions used to model the spectra of FK Aqr and FL Aqr. The posterior distributions of the shape and rate parameters are shown in the left and right panels, respectively. . . . .	95
2.11	Chandra observation of a crowded field near the center of the Orion Nebula Cluster. This field is approximately $25'' \times 25''$ in size, and is centered at (RA,Dec)=(5:35:15.4,-05:23:04.68). Shown in blue are approximate 90% posterior credible regions for source locations, under the spatial-only model (left), and the extended full model (right). The figures next to the regions indicate the estimated relative intensities. The credible region of the source with the largest location uncertainty is circled in green (right panel). The red rectangular box encloses two overlapping sources (right panel) for which we carry out a detailed follow-up spectral analysis (Section 2.7.2). . . . .	98

2.12	Number of sources detected in the analysis of the Chandra observation in Figure 2.11. Posterior of $K$ based on the spatial-only model (left) and the extended full model (right). . . . .	98
2.13	Detailed spectral analysis of overlapping COUP sources #732 and #744. Best-fit values of absorption column ((a), (b)), temperature ((c), (d)), metallicity ((e), (f)), and flux ((g), (h)) for the disentangled analysis, for each of 1000 allocations of the photons are shown as histograms. Panels (a), (c), (e), and (g) correspond to the bright source and panels (b), (d), (f), and (h) correspond to the fainter source. The naïve analysis best-fit values and their 68% intervals are shown by the solid and dashed red vertical lines, respectively. The width of the histograms only account for uncertainty due to the allocation of photons, and not additional statistical error, which is well described by the intervals shown for the naïve analysis. . .	103
3.1	Orbit of $(x, 1)$ under the invariance group $\mathcal{G}_{W_3}$ that corresponds to Warp III transformations. . . . .	134
3.2	Illustration of a sample augmentation that facilitate multiple cross space transformations in a group invariant sub-model. . . . .	135
3.3	Plots of the $N(10, 2)$ and Skew-Normal(0, 3, 10) probability densities used in our simulation study. . . . .	146
3.4	Estimates of $r$ across the 500 simulations; the rows correspond to the three methods of choosing $D$ (described in the main text) and the columns correspond to the two simulation settings $n_1 = n_2 = 20$ (left) and $n_1 = n_2 = 100$ (right). . . . .	147
A.1	Trace plot of the parameter $\mu_{5x}$ from a simulation with ten sources (Section 2.5.1) before (left) and after (right) relabelling. . . . .	160
A.2	2-D King profile density (left), and its contours (right). . . . .	161

# List of tables

- 2.1 Symbols used in this chapter. Notation used only in a single section is defined where it appears and is not included in this table. . . . 50
- 2.2 Fitted parameters under the full and spatial-only models. The columns in bold give the fits that would likely be relied upon in practice for the two models. The intervals in parentheses indicate the 16% and 84% posterior quantiles, i.e., Bayesian  $1\sigma$  equivalent intervals. . . . . 67
- 2.3 Photon allocation proportions for the spatial-only and full models. 67
- 2.4 Posterior means under the spatial-only model and the full model. The parenthetic intervals are  $1\sigma$  error bars computed using 16% and 84% posterior quantiles. . . . . 94
- 2.5 Extended full model fit for the Chandra observation in Figure 2.11. Posterior mean locations and relative intensities (as percentages), with 68% intervals indicated. . . . . 99
  
- 3.1 Cayley table of the invariance group  $\mathcal{G}_{W3}$  used by the sub-model of the measure that corresponds to Warp III transformations. . . . . 133

# Acknowledgments

First and foremost, I thank my advisor Xiao-Li Meng for his terrific guidance throughout my graduate studies. He has taught me a great deal of the art and science of statistical research including the power of deep questions, the reach of statistical principles, and the importance of good writing and memorable presentations. I especially thank Xiao-Li for his many insights that contributed to this thesis and his patience and generosity in supporting me. Despite his demanding role as Dean of the Graduate School of Arts and Sciences, Xiao-Li has always set aside time to meet with me, comment on my work, and offer his wisdom. I deeply appreciate all of his guidance and I am sure that his teaching and inspiration will stay with me throughout my career and life.

I am also greatly indebted to David van Dyk and Vinay Kashyap who have both been excellent inspirations to me and have made me feel welcome in the exciting world of astrostatistics. Their suggestions and extensive comments very much improved Chapter 2 of this thesis, and my collaboration with them has been a key part of my studies. I also thank all the members of the CHASC International Astrostatistics Center; they have been wonderful to work with and I greatly appreciate all the discussions we had and the feedback I received over the years.

I am sincerely grateful to Tirthankar Dasputga for being on my thesis committee and for generously giving his time to support me in research and career

development. His comments on the material in Chapter 1 were particularly helpful, and it would be an honor to work more with him in the future.

All of the faculty in the Statistics Department have made splendid contributions to my education, and I especially thank those not yet mentioned whose courses have greatly deepened my understanding of statistics, namely Joseph Blitzsten, Carl Morris, Jun Liu, Natesh Pillai, Donald Rubin, and Samuel Kou. I also thank all the department staff for their magnificent help, especially Betsey Cogswell, Madeleine Straubel, James Matejek, Alice Moses, and Maureen Stanton.

It is my great pleasure to acknowledge many of my fellow graduate students in the Statistics Department who have added to my experience at Harvard in innumerable ways and without whom this thesis would scarcely have been possible. I thank Arman Sabbaghi, Lazhi Wang, David Watson, Ruobin Gong, Yang Chen, Nathan Stein, Dan Cervone, Yang Li, Alex Blocker, Hyungsuk Tak, Viviana Garcia, Peng Ding, Lo-Hua Yuan, Xufei Wang, Sobambo Sosina, and many others. Special thanks go to Arman Sabbaghi and Lazhi Wang for all their help with preparation for the qualifying examination, research ideas, and my future path.

Most importantly, I express my great gratitude to my parents, my siblings, and my loyal friends Jing Ma, Christopher Miller, and Anna Wang. They have all been terrifically supportive during my studies, and it is to them that I owe my deepest enthusiasm for learning and life.

Lastly, I appreciate the funding I have received from NSF and the Smithsonian Competitive Grants Program for Science while conducting this research at Harvard University.





To my parents.

# 1

## Designing test information and test information in design

### 1.1 Motivation and overview

#### 1.1.1 Test information: foundations and developments

Nicolae et al. (2008) highlight that the amount of information provided by an experiment depends on our goals, and in particular the amount of information for hypothesis testing can be different to that for estimation. Nonetheless, the

importance of information measures and the need for a framework for constructing and understanding them is common to both the testing and estimation scenarios. Indeed, [Ginebra \(2007\)](#) points out that flexible information measures are essential because information is a “highly multi-dimensional concept” which cannot be captured by a narrow definition.

In statistics, perhaps the most famous appeal to information is in likelihood theory which gives the asymptotic variance of the maximum likelihood estimator (MLE) as the inverse of the Fisher information. But, the key to the importance of information measures is that they quantify what it is possible to learn on average given a data generating model (for the data to be used), and thus they go beyond detailing the properties of a specific procedure. This is illustrated by the fact that the Fisher information is not *merely* related to the asymptotic variance of the MLE, it also appears in the Cramér-Rao lower bound for the variance of all unbiased estimators. Given such requirements, it is natural that in the Bayesian case estimation information measures are based on the posterior distribution, because it contains all that can be known about the parameters. Furthermore, it should be no surprise that the fundamental component of our test information measures is the Bayes factor or likelihood ratio.

Development of the existing estimation information framework began when [Shannon \(1948\)](#) gave his now well-known definition of statistical entropy,

$$H(\pi) = E_{\theta}[-\log \pi(\theta)] = - \int_{\Theta} \pi(\theta) \log \pi(\theta) \mu(d\theta). \quad (1.1)$$

This is a measure of the information gained by observing the random variable  $\theta \in$

$\Theta$  with density  $\pi$ , with respect to the measure  $\mu$ , which is typically the Lebesgue or counting measure. Next, Lindley (1956) defined the expected information about a parameter provided by an experiment as the difference between the prior entropy and the expected posterior entropy. This measure leads to the D-optimality design criterion that is now often used in Bayesian experimental design, see for example the review by Chaloner and Verdinelli (1995). It has also been used for other purposes, such as defining reference priors, an example being the work of Berger et al. (2009). A framework was proposed by DeGroot (1962) which generalizes Lindley’s definition of the information provided by an experiment to the difference between the prior and the expected posterior *uncertainty*, where the uncertainty need not be quantified by entropy. Furthermore, DeGroot (1962) demonstrated that many of Lindley’s results do not rely on the specific mathematical form of entropy, and therefore they carry over to his more general information measures.

Our first theoretical contribution is to synthesize the test information measures suggested by Nicolae et al. (2008) to develop an analogous framework to that of DeGroot (1962) for test information. The general measures of expected test information that we propose use the  $f$ -divergence introduced by Csiszár (1963) and Ali and Silvey (1966), and we extend the concept to define observed and conditional test information because these are of great importance in sequential design. The use of divergence measures is natural and highlights fundamental differences between testing and estimation scenarios, namely that in the testing case the two hypotheses to be compared must be incorporated, and in computing expected test information we must choose one of the hypotheses to condition

on. This last observation suggests that every expected test information measure should have a dual which conditions on the other hypothesis. Furthermore, we may anticipate that there is an appealing subclass of measures which give the same expected test information as their duals, an insight that we will revisit.

Our second theoretical contribution is to establish further connections between test and estimation information measures. These connections concern an important quantity for sequential design discussed by [Nicolae and Kong \(2004\)](#) and [Nicolae et al. \(2008\)](#), namely the fraction of information contained in the observed data relative to that contained in the complete data (which additionally includes unobserved or missing data). [Nicolae et al. \(2008\)](#) established an asymptotic equivalence between their measures of the fraction of observed test information and the fraction of observed Fisher information (for estimation), as the distance between the null parameter and the MLE goes to zero. Intuitively, this means that the fraction of observed Fisher information is a good approximation to the fraction of observed test information if the null is close to the truth. We show that, by allowing different weighting of observed and missing Fisher information, the equivalence can be extended to hold for all test information measures under our framework. This result further links test and estimation information and identifies an appealing class of test information measures that weight observed and missing *estimation* information equally (in the limit considered).

With the basic foundations of our test information framework in place, we consider its practical implications. [Nicolae and Kong \(2004\)](#) and [Nicolae et al. \(2008\)](#) put forward their measures of the fraction of observed (or missing) test information

with the purpose of informing data collection decisions in genetic linkage studies. We now build on this by illustrating specifically how test information measures may be used in experimental design, both in model selection and coefficient testing scenarios. In the design for model selection scenario, it is often not clear how to use estimation information measures, but the use of test information measures is intuitive. We demonstrate this by finding optimal designs for choosing between the complementary log-log and Probit link functions for binary linear regression.

Next, in the case of testing for linear regression parameters, we give a closed form design criterion that is related to the familiar Bayesian alphabet optimality criteria used in estimation contexts, though it is specific to testing. Not surprisingly, in our simulations, optimizing this criterion yields designs which perform better than the commonly used D-optimality criterion, for the purpose of testing. We also propose a posterior probability based expected test information measure, which has many appealing properties, and similarly outperforms D-optimal designs. In fact, this measure gives the same values as its dual and therefore exhibits the property we have anticipated. Consequently, the design optimized with respect to this measure does not depend on which hypothesis is true for its optimality, a particularly useful symmetry in practice. Also in the linear regression coefficient testing scenario, [Toman \(1996\)](#) showed that, for a particular loss function, minimizing the Bayes risk corresponds to choosing the D-optimal design. This approach and conclusion differs to ours since we maximize the expected probabilistic information for distinguishing hypotheses, rather than minimizing an expected loss.

Much of the other experimental design literature focuses only on estimation problems, but among the limited work that deals with design for testing and model selection, the approach of [Box and Hill \(1967\)](#) is perhaps most similar to ours. They choose designs that maximize the entropy of the posterior probability mass function of the model indicator, but do not offer a flexible framework for test information measures and design, as we do. In terms of mathematical justification, our framework benefits from the work of [Ginebra \(2007\)](#), which reviews and synthesizes previous theory including that of [Blackwell \(1951\)](#), [Sherman \(1951\)](#), [Stein \(1951\)](#), and [Le Cam \(1964\)](#). Specifically, our expected test information measures satisfy (up to aesthetics) the three basic requirements that [Ginebra \(2007\)](#) argues are essential for any expected information measure, regardless of the context. But our framework also adds fresh perspectives, indeed the concept of coherent dual test information measures is new and fundamental, and our observed test information measures have fewer restrictions than those suggested by [Ginebra \(2007\)](#) (who focused on the estimation case). Furthermore, we emphasize that in practice test and estimation information measures behave very differently, despite their common mathematical roots described by [Ginebra \(2007\)](#). This fact does not seem to have been discussed in detail before.

A key limitation of designs optimized for distinguishing between a null and an alternative model (or set of models) is their inherent sensitivity to these hypothesized models. In particular, if none of the hypothesized models are a close approximation to the unknown true data generating model, then the observed data are unlikely to be as we predict, and consequently the design we select may be

sub-optimal for choosing between our hypotheses. This issue is unavoidable and can only be mitigated by generic space filling designs sometimes used in estimation scenarios for similar reasons. Despite this limitation, test information measures are valuable design tools because many scientific investigations are of confirmatory nature, meaning that there is some reason to believe that the proposed models capture scientifically acceptable descriptions (this particularly holds in physics). In summary, the information measures we propose are beneficial whenever an investigator seeks to compare several reasonable competing models.

This chapter is organized as follows. The remainder of Section 1.1 gives three categories of scientific problem where test information measures are useful, briefly reviews the estimation information framework proposed by [DeGroot \(1962\)](#), and discusses the parallels with the test information measures introduced by [Nicolae et al. \(2008\)](#). The main body of the paper is divided between Sections 1.2 and 1.3, which deal with expected and observed test information, respectively. These sections finish with illustrations of the practical use of test information in design and sequential design for decision problems, respectively. Section 1.2 also introduces a fundamental symmetry condition that defines an appealing class of test information measures. Section 1.4 presents our main result linking test and estimation information, and Section 1.5 concludes with discussion and open problems.

### **1.1.2 Uses of test information**

We now describe three experimental design problems, representing broad categories of scientific questions (shown in *italics*), for which test information measures



are useful.

*Classification and model selection.* In astronomy, the intensity of some sources (e.g., Cepheid stars) varies periodically over time, thus creating a continuous function of time called a lightcurve. Such sources can be classified by features of their lightcurves, e.g., the period. Since telescope time is limited for any group of researchers, the lightcurve of a source is observed at a number of time points and then a classifier is applied. For example, some modern techniques use random forest classifiers, e.g., [Richards et al. \(2011\)](#) and [Dubath et al. \(2011\)](#). Intuitively, given some lightcurve observations, the design problem is to pick the time of the next observation that will maximize the probability of correct classification.

*Screening and follow-up.* In genetic linkage studies it is of interest to test if markers (or genes) located close together on the same chromosome are more likely to be inherited together than if markers are inherited independently (the null hypothesis). This is a screening process because the magnitude of the linkage (i.e., dependence) is ultimately of interest. In the case of too much missing information, a follow-up study could choose between increasing the number of markers in potential regions of linkage or increasing the sample size (the number of people). To assess which option is likely to have greater power, for example, we must take the models under the two hypotheses into account. See [Nicolae and Kong \(2004\)](#) and [Nicolae et al. \(2008\)](#) for previous work on this problem.

*Robust design.* Test information measures can also be useful in applications at the interface of testing and estimation. In chemical engineering, it is often of interest to estimate the mean yield of a product under different conditions, and

ultimately to model the yield. In this situation, space filling designs are usually preferred because it is unknown where the regions of high (and low) yield will be. However, space filling designs can vary in their efficiency for distinguishing specific models, and test information measures can be used to select the ones that best separate important candidate models.

### 1.1.3 Bayesian information for estimation

We briefly review the framework of DeGroot (1962) to help make clear both distinctions and parallels between test and estimation information. Suppose that we are interested in a parameter  $\theta \in \Theta$  and our prior distribution is  $\pi$ , where  $\Theta$  is the parameter space. Information about the parameter is gained through an experiment  $\xi$  whose future outcome is the random variable  $X \in \mathcal{X}$ , where  $\mathcal{X}$  denotes the set of possible outcomes of the experiment. For example,  $\xi$  may specify the design points at which data are to be collected. We denote by  $\mathcal{I}(\xi; \pi)$  a measure of the expected information to be gained by conducting  $\xi$ .

The measure  $\mathcal{I}(\xi; \pi)$  should have basic properties such as non-negativity and additivity. To specify the meaning of additivity we need the notion of conditional information: if  $\xi = (\xi_1, \xi_2)$  is an experiment composed of two sub-experiments, then we denote by  $\mathcal{I}(\xi_2|\xi_1; \pi)$  the expected *conditional* information to be gained by conducting  $\xi_2$  *after* conducting  $\xi_1$ , i.e., the expected new information that will be gained from  $\xi_2$ . Additivity can now be formalized.

**Definition 1.1** *An information measure  $\mathcal{I}$  is additive if, for any composite ex-*

periment  $\xi = (\xi_1, \xi_2)$  and any proper prior  $\pi$ , the following relation holds

$$\mathcal{I}(\xi; \pi) = \mathcal{I}(\xi_1; \pi) + \mathcal{I}(\xi_2|\xi_1; \pi). \quad (1.2)$$

DeGroot (1962) chooses  $\mathcal{I}(\xi; \pi)$  to be the difference between the prior uncertainty and the expected posterior uncertainty about  $\theta$ . In particular, he defines the prior uncertainty to be  $U(\pi)$ , where the *uncertainty function*  $U$  is a concave functional of  $\pi$ , i.e.,  $U(\lambda\pi_1 + (1 - \lambda)\pi_2) \geq \lambda U(\pi_1) + (1 - \lambda)U(\pi_2)$  for any two densities  $\pi_1$  and  $\pi_2$  and  $\lambda \in [0, 1]$ . Similarly, DeGroot (1962) defines the expected posterior uncertainty to be  $E_X[U(p(\cdot|X))]$ , where the expectation is with respect to  $f(x) = \int_{\Theta} f(x|\theta)\pi(\theta)\mu(d\theta)$ . Thus, we have the following.

**Definition 1.2** *The expected Bayesian estimation information provided by an experiment  $\xi$ , under a proper prior  $\pi$ , is*

$$\mathcal{I}(\xi; \pi) = U(\pi) - E_X[U(p(\cdot|X))]. \quad (1.3)$$

Lindley (1956) suggests  $U$  should be the entropy function  $H$  given in (1.1). DeGroot (1962) shows that (1.3) satisfies non-negativity for all priors  $\pi$  and experiments  $\xi$  if and only if  $U$  is concave. To generalize further, we follow the logic of Definition 1.2 and define the expected conditional estimation information contained in the second of two sub-experiments as

$$\mathcal{I}(\xi_2|\xi_1; \pi) = E_{X_1}[U(p(\cdot|X_1))] - E_{X_1, X_2}[U(p(\cdot|X_1, X_2))], \quad (1.4)$$

where the second expectation is with respect to the joint density of  $X_1$  and  $X_2$ , with  $X_1$  and  $X_2$  being the outcomes of  $\xi_1$  and  $\xi_2$ , respectively. The fact that the measure given in Definition 1.2 is additive now follows directly from adding (1.3) (with  $X_1$  replacing  $X$ ) and (1.4), because  $E_{X_1}[U(p(\cdot|X_1))]$  cancels. However, it is not generally true that  $\mathcal{I}(\xi; \pi) = \mathcal{I}(\xi_1; \pi) + \mathcal{I}(\xi_2; \pi)$ , even under independence of  $X_1$  and  $X_2$ .

#### 1.1.4 Insights from test information measures proposed by Nicolae et al. (2008)

For the sharp test hypotheses  $H_0 : \theta = \theta_0$  and  $H_1 : \theta = \theta_1$ , Nicolae et al. (2008) (implicitly) propose the very natural frequentist expected test information measure

$$\mathcal{I}^T(\xi; \theta_0, \theta_1) = E_X[\log \text{LR}(\theta_1, \theta_0|X)|\theta_1], \quad (1.5)$$

where the superscript  $T$  indicates the testing context, and

$$\text{LR}(\theta_0, \theta_1|x) = \frac{f(x|\theta_0)}{f(x|\theta_1)} \quad (1.6)$$

is the likelihood ratio. We observe that (1.5) is the Kullback-Leibler (KL) divergence between the data models  $f(\cdot|\theta_0)$  and  $f(\cdot|\theta_1)$ , and thus it is closely connected to the entropy based measure proposed by Lindley (1956). (The KL divergence between two densities  $g$  and  $h$ , will be denoted  $KL(g||h)$ , and is defined

as  $\int_{\mathcal{X}} g(x) \log(g(x)/h(x)) \mu(dx)$ , where the support of  $g$  is assumed to be a subset of the support of  $h$ .) Nonetheless, there is a good reason why [Nicolae et al. \(2008\)](#) did not simply use Definition 1.2 to construct measures of test information; namely, it does not take the hypotheses into account. Indeed, the presence of the two parameter values,  $\theta_0$  and  $\theta_1$ , in (1.5) clearly distinguishes test information from the estimation information we have considered so far. This difference makes intuitive sense because it represents the difference between gaining evidence for distinguishing two hypotheses and neutrally gaining knowledge about the parameter.

In practice, the alternative hypothesis is often composite and in the Bayesian context we then write  $H_1 : \theta \sim \pi$ , for some prior  $\pi$ . One of the Bayesian measures of expected test information (implicitly) suggested by [Nicolae et al. \(2008\)](#) is

$$\mathcal{I}^T(\xi; \theta_0, \pi) = \text{Var}_{\theta, X}(\text{LR}(\theta_0, \theta|X)). \quad (1.7)$$

Variance and entropy are both measures of spread and hence (1.7) is also connected to the measure proposed by [Lindley \(1956\)](#), although no logarithm is taken in (1.7). The key distinction with Definition 1.2 is again due to the appearance of the null hypothesis  $\theta_0$ . In summary, these examples have connections with the estimation information measures reviewed in Section 1.1.3, but also have common features distinguishing test information from estimation information. Based on these parallels and distinctions, the next section proposes our general framework for constructing test information measures.

## 1.2 Expected test information: general framework and applications

### 1.2.1 Test information: a synthesis

The two key properties of expected information measures are non-negativity and additivity. For simplicity, we develop our framework in the case of continuous densities and the Lebesgue measure. Theorem 2.1 of DeGroot (1962) establishes non-negativeness of the estimation information reviewed in Section 1.1.3. Writing the marginal density of  $x$  as  $f(x) = \int_{\Theta} f(x|\theta)\pi(\theta)d\theta$ , the key equation underlying the theorem is

$$E_X[p(\cdot|X)] = \int_{\mathcal{X}} p(\cdot|x)f(x)dx = \pi(\cdot). \quad (1.8)$$

That is, the expected posterior density with respect to the marginal density is the prior density. To see the corresponding key identity for hypothesis testing, we first observe that the expected test information (1.5) uses the likelihood ratio as the fundamental statistic for quantifying the information for distinguishing two values of  $\theta$ . More generally, the hypotheses may be composite, say  $H_0 : \theta \in \Theta_0$  and  $H_1 : \theta \in \Theta_1$ , in which case we turn to the Bayesian perspective and replace the likelihood ratio with the Bayes factor

$$\text{BF}(x|H_0, H_1) = \frac{f(x|H_0)}{f(x|H_1)} = \frac{\int_{\Theta_0} f(x|\theta)\pi(\theta|H_0)d\theta}{\int_{\Theta_1} f(x|\theta)\pi(\theta|H_1)d\theta} = \frac{\int_{\Theta_0} f(x|\theta)\frac{\pi(\theta)}{\pi_0}d\theta}{\int_{\Theta_1} f(x|\theta)\frac{\pi(\theta)}{\pi_1}d\theta}, \quad (1.9)$$

where  $\pi_i = P(\theta \in \Theta_i)$ , for  $i = 0, 1$ . (We assume  $\pi_i \neq 0$  throughout, for  $i = 0, 1$ .)

Thus, for hypothesis testing, the analogous equation to (1.8) is

$$E_X[\text{BF}(X|H_0, H_1)|H_1] = \int_{\mathcal{X}} \frac{f(x|H_0)}{f(x|H_1)} f(x|H_1) dx = 1. \quad (1.10)$$

That is, the expected Bayes factor (or likelihood ratio), under the alternative, does not favor either hypothesis. For simplicity, we assume here and throughout that the support of  $f(\cdot|\theta)$  is  $\mathcal{X}$  for all  $\theta \in \Theta$ . Equation (1.10) allows us to apply Jensen's inequality to ensure that the general expected test information given in Definition 1.3 (below) is non-negative. For test information, the parallel of the uncertainty function  $U$  is the *evidence function*  $\mathcal{V}$ , which acts on the positive real numbers and in particular has the Bayes factor (or likelihood ratio) as its argument. The use of Jensen's inequality to ensure non-negativity requires that the evidence function is concave, and we therefore assume concavity throughout. Note that, what is measured by the evidence function is the evidence *in support of the null*, and therefore, like DeGroot (1962), we are interested in a reduction.

**Definition 1.3** Under  $H_1 : \theta \in \Theta_1$ , the expected test information provided by the experiment  $\xi$  for comparing the hypotheses  $H_0 : \theta \in \Theta_0$  and  $H_1 : \theta \in \Theta_1$ , for a given evidence function  $\mathcal{V}$  and a proper prior  $\pi$ , is defined as

$$\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi) = \mathcal{V}(1) - E_X[\mathcal{V}(\text{BF}(X|H_0, H_1))|H_1], \quad (1.11)$$

where  $\Theta_0 \cap \Theta_1 = \emptyset$ .

The prefix “expected” is necessary because the Bayesian approach generally assumes that data have been observed. Note that, (1.11) is mathematically equivalent to the frequentist measure

$$\mathcal{V}(1) - E_X[\mathcal{V}(\text{LR}(\theta_0, \theta_1|X)|\theta_1)] \tag{1.12}$$

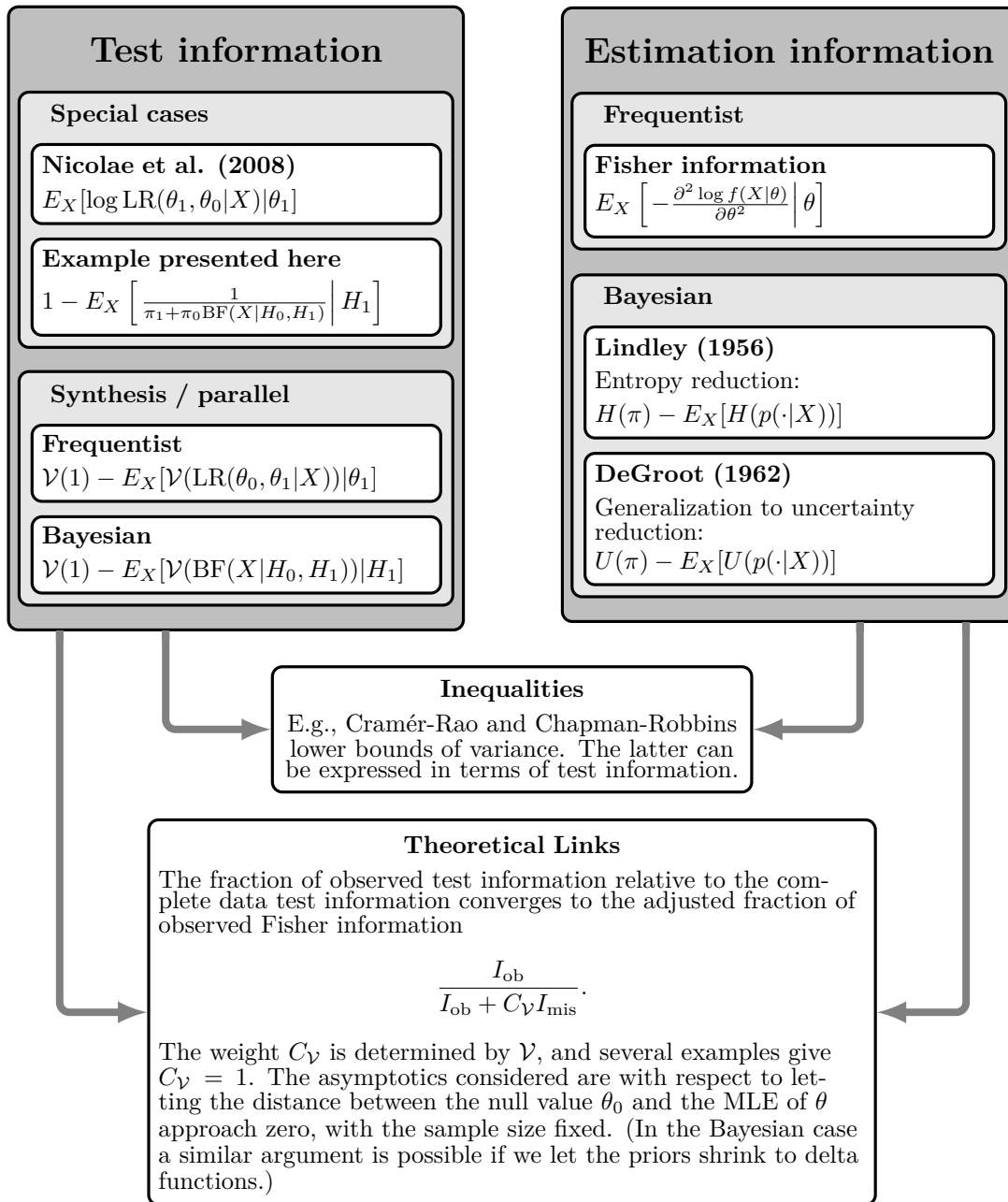
when  $\Theta_i = \{\theta_i\}$  and  $P(\theta = \theta_i) \neq 0$ , for  $i = 0, 1$ . (The frequentist perspective is also recovered if the prior is viewed as part of the data generating model.) Under (1.12), the measure (1.5) is given by choosing  $\mathcal{V}(z) = \log(z)$ . The mathematical equivalence of Bayesian and frequentist measures of expected test information means that we can interchange the Bayes factor in (1.11) and the likelihood ratio as convenient. More generally, the Bayes factor in (1.11) can be replaced by *any* numerical comparison of the hypotheses, at least if the baseline is also adjusted. However, the main focus here will be on the Bayesian perspective because it is statistically coherent and is conceptually well suited to incorporating composite hypotheses (and nuisance parameters, see Section 1.2.3) when no data have been observed, as is often the case when we choose a design. We retain the argument  $\pi$  in our notation  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi)$  as a reminder that (1.11) does depend on the prior  $\pi$ , which we should therefore choose careful, as with the specification of any part of our models. Also note that, the parameter in Definition 1.3 can simply be a model indicator and hence our framework goes beyond parametric models.

The final term of (1.11) is the  $f$ -divergence introduced by [Csiszár \(1963\)](#) and [Ali and Silvey \(1966\)](#), which generalizes KL divergence. Indeed, as mentioned in Section 1.1.4, the measure (1.5) is a KL divergence. The properties of KL



divergence alert us to the important feature that expected test information is not necessarily symmetric in the two hypotheses. A class of evidence function that treat the hypotheses equally will be introduced in Section 1.2.2.

The baseline term  $\mathcal{V}(1)$  does not appear in (1.5) because it turns out to be zero, but in general  $\mathcal{V}(1)$  ensures non-negativity of expected test information, and it is also one of the ingredients needed for (1.11) to be interpreted as a general information measure, as we now explain. A general information measure should appeal to many different researchers and this is typically achieved by considering *maximal* information. For example, Fisher information measures the maximal estimation information asymptotically available. Our test information measures and Degroot’s estimation information measures (see Definition 1.2) are also implicitly maximal since all relevant information is contained in the Bayes factor (or likelihood ratio) and the posterior distribution, respectively. For testing, researchers must choose a test statistic (not necessarily the Bayes factor or likelihood ratio) and quantities such as the size of the test (in frequentist settings), but these decisions do not affect the maximal test information available. Thus, the KL divergence in (1.5) quantifies how well the hypotheses can be separated probabilistically, but the decision as to whether to use data in a statistically efficient way and how to interpret the observed separation (e.g., whether to reject the null) is left to the individual investigators. In general, if the baseline value in (1.11) depended on investigator specific decisions, then the notion of maximal information would be lost. Fortunately, the choice of  $\mathcal{V}(1)$  as the baseline value does not suffer from this problem, and has common appeal because it represents no evidence for



**Figure 1.1:** Summary of estimation and test information theory. The synthesis of test information measures into one coherent framework paralleling the estimation framework is new. Also new are the general links between estimation and test information, although Nicolae et al. (2008) considered the same connection with Fisher information for specific cases.

either hypothesis.

The test information introduced above and the links and parallels with existing estimation information measures are summarized in Figure 1.1 (the theoretical links are discussed in detail in Section 1.4). From this point on, we will frequently write  $\mathcal{I}_{\mathcal{V}}^T(\xi)$  to mean  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi)$  and for similar notation will again drop the arguments after the ‘;’ symbol when this causes no confusion. We complete our initial development of test information by specifying the form of  $\mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1; H_0, H_1, \pi)$ , which is easily deduced from the expected conditional estimation information (1.4).

**Definition 1.4** *The expected conditional test information  $\mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1)$  provided by conducting the experiment  $\xi_2$  after  $\xi_1$  is*

$$\mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1; H_0, H_1, \pi) = E_{X_1}[W(X_1)|H_1] - E_{X_1, X_2}[W(X_1, X_2)|H_1], \quad (1.13)$$

where  $W(X_1) = \mathcal{V}(\text{BF}(X_1|H_0, H_1))$  and  $W(X_1, X_2) = \mathcal{V}(\text{BF}(X_1, X_2|H_0, H_1))$ .

That (1.13) is non-negative is again a consequence of Jensen’s inequality;

$$E_{X_2}[W(x_1, X_2)|H_1, x_1] \leq W(x_1), \quad (1.14)$$

where, to make the expressions easier to read, we have denoted the observed data by lower case letters, and unobserved data by upper case letters. Given Definition 1.4, the additivity property of Definition 1.1 holds trivially, i.e.,  $\mathcal{I}_{\mathcal{V}}^T(\xi) = \mathcal{I}_{\mathcal{V}}^T(\xi_1) + \mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1)$ .

### 1.2.2 Symmetry class and a probability based measure

The best choice of  $\mathcal{V}$  will to some extent depend on the particular context (see Section 1.4 for some theoretical guidance), but here we propose a class of evidence functions that have appealing properties. The class is those evidence functions that treat the hypotheses symmetrically and in particular satisfy the condition

$$\frac{\mathcal{V}(z; H_0, H_1)}{\mathcal{V}(1/z; H_1, H_0)} = z. \quad (1.15)$$

Naturally,  $\mathcal{V}(1/z; H_1, H_0)$  represents the evidence for the alternative, since the roles of  $H_0$  and  $H_1$  have been swapped. Thus, setting  $z = \text{BF}(x|H_0, H_1)$ , the symmetry condition (1.15) states that our choice of  $\mathcal{V}$  should preserve the Bayes factor. We include the arguments  $H_0$  and  $H_1$  in (1.15) because in general the evidence measures may be allowed to depend on the order of the hypotheses through prior probabilities as well as through the Bayes factor.

The symmetry condition (1.15) has the important consequence that the resulting expected test information measure satisfies the fundamental coherence identity

$$\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi) = \mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0, \pi). \quad (1.16)$$

The right hand of (1.16) swaps the hypotheses, indicating that the *dual* expected test information measure  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0, \pi)$  takes an expectation with respect to  $f(\cdot|H_0)$ , rather than  $f(\cdot|H_1)$ . Indeed, the dual test information measure quantifies the reduction in evidence for the alternative when data are collected under the null. The coherence identity (1.16) states that, before we observe any data, the

expected amount of information in the data for distinguishing the two hypotheses is the same regardless of which is in fact true. This symmetry is intuitive because the probabilistic separation of the two marginal data models  $f(\cdot|H_0)$  and  $f(\cdot|H_1)$  does not depend on which hypothesis is true. The practical importance of the coherence identity is that, when optimizing the expected test information with respect to the experiment  $\xi$ , we do not need to know which hypothesis is true to ensure the optimality of our experiment.

We can go further and consider what specific evidence functions satisfying (1.15) are particularly appealing. We want our evidence function to be probability based because hypothesis testing is fundamentally about seeking probabilistic evidence, usually in the form of p-values or posterior probabilities. Indeed, for the purposes of test information, the traditional estimation information focus on variance and spread is in general inadequate. From the Bayesian perspective, a sensible probability based evidence function is

$$\mathcal{V}(z; H_0, H_1) = \frac{z}{\pi_1 + \pi_0 z}, \quad (1.17)$$

where  $\pi_0$  and  $\pi_1$  are the prior probabilities of  $H_0$  and  $H_1$ , respectively (for simplicity we assume  $\pi_0 + \pi_1 = 1$ ). When  $z = \text{BF}(x|H_0, H_1)$ , then (1.17) is just the posterior to prior probability ratio for  $H_0$ . The resulting dual expected test

information measures are

$$\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1) = 1 - E_X \left[ \frac{z(X)}{\pi_1 + \pi_0 z(X)} \middle| H_1 \right] \quad (1.18)$$

$$\mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0) = 1 - E_X \left[ \frac{1}{\pi_1 + \pi_0 z(X)} \middle| H_0 \right], \quad (1.19)$$

where  $z(X)$  denotes the Bayes factor  $\text{BF}(X|H_0, H_1)$ . The measure (1.18) is simply the expected difference between the prior and posterior probability of the null, relative to the prior probability, when the data are actually from the alternative. That is, the relative loss in probability of the null. The measure (1.19) is the same but with the roles of  $H_0$  and  $H_1$  switched. Since (1.15) is satisfied, the coherence identity (1.16) tells us that  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1) = \mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0)$ . This and the straightforward Bayesian probability interpretation of (1.18) make (1.17) a particularly appealing choice of evidence function.

There are also other examples of evidence functions that satisfy (1.15), e.g.,

$$\mathcal{V}(z) = \frac{1}{2} \log(z) - \frac{1}{2} z \log(z). \quad (1.20)$$

For this evidence function, both sides of (1.16) equal  $\frac{1}{2}KL(f(\cdot|H_1)||f(\cdot|H_0)) + \frac{1}{2}KL(f(\cdot|H_0)||f(\cdot|H_1))$ . Historically, this symmetrized form of KL divergence is the divergence that [Kullback and Leibler \(1951\)](#) originally suggested (without scaling by a half). Intuitively, it can be interpreted as a measure of the expected test information when the two hypotheses are considered equally likely apriori. However, symmetrized KL divergence does not have a straightforward probability interpretation, and therefore we prefer (1.17)-(1.19).

### 1.2.3 Nuisance parameters

Many statistical problems come with nuisance parameters. In the frequentist setting, once data have been observed, estimates of the nuisance parameters can be plugged in to give a point estimate of the expected test information (1.12) (as can estimates of  $\theta_0$  and  $\theta_1$  when the hypotheses are composite). A confidence interval for (1.12) can be obtained by evaluating it for values of the nuisance parameters within a confidence interval. (Both could instead be done for observed test information, see Section 1.3.1.) In design problems, data are typically yet to be collected but (1.12) could be evaluated on a grid of values of the nuisance parameters.

In the Bayesian context, the nuisance parameters  $\beta_0$  (under the null) and  $\beta_1$  (under the alternative) are simply integrated out along with the parameters that define the hypotheses. That is,

$$\mathcal{I}_V^T(\xi; H_0, H_1, \pi, \psi_0, \psi_1) = \mathcal{V}(1) - E_X [\mathcal{V}(\text{BF}(X|H_0, H_1)) | H_1], \quad (1.21)$$

where the Bayes factor is now given by

$$\frac{\int_{\Theta_0} \int_{B_0} f(X|\theta, \beta_0) \psi_0(\beta_0|\theta) \pi(\theta|H_0) d\beta_0 d\theta}{\int_{\Theta_1} \int_{B_1} f(X|\theta, \beta_1) \psi_1(\beta_1|\theta) \pi(\theta|H_1) d\beta_1 d\theta}, \quad (1.22)$$

with  $B_i$  being the support of the prior density  $\psi_i$  of  $\beta_i$ , for  $i = 0, 1$ . Clearly, the mathematical properties of (1.21) are the same as those of (1.11).

As mentioned in Section 1.2.1, alternatives to the Bayes factor in Definition 1.3 can be used at the expense of the coherence of the Bayesian method and simplicity.

For example, those intending to use the likelihood ratio test, may opt to mimic the likelihood ratio test statistic by calculating

$$\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi, \psi_1) = \mathcal{V}(1) - E_X \left[ \mathcal{V} \left( \frac{f(X|\theta_{\text{MLE}}^{H_0}, \beta_{0,\text{MLE}})}{f(X|\theta_{\text{MLE}}^{H_1}, \beta_{1,\text{MLE}})} \right) \middle| H_1 \right], \quad (1.23)$$

where  $\theta_{\text{MLE}}^{H_i}$  and  $\beta_{i,\text{MLE}}$  are the MLEs of  $\theta$  and  $\beta_i$ , respectively, under hypothesis  $H_i$ , for  $i = 0, 1$ . In this work we focus on the expected test information given in Definition 1.3 (and (1.12)) and thus leave the theoretical investigation of measures such as (1.23) for future work. However, we include numerical results based on (1.23) in Section 1.2.4.

### 1.2.4 Binary regression example: distinguishing the complementary log-log and Probit link functions

Consider the binary regression model

$$X_i | M, \beta_\theta, g_\theta \sim \text{Bernoulli}(p_i), \quad (1.24)$$

for  $i = 1, \dots, n$ , where

$$M^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \end{pmatrix} \quad (1.25)$$

is the design matrix (i.e., essentially  $\xi$ ), and  $g_\theta(p_i) = \beta_{\theta,\text{int}} + \beta_{\theta,\text{slope}} t_i$ , for the link function  $g_\theta$  and  $\theta \in \{0, 1\}$ . The sharp hypotheses of interest are  $H_0 : \theta = 0$  and

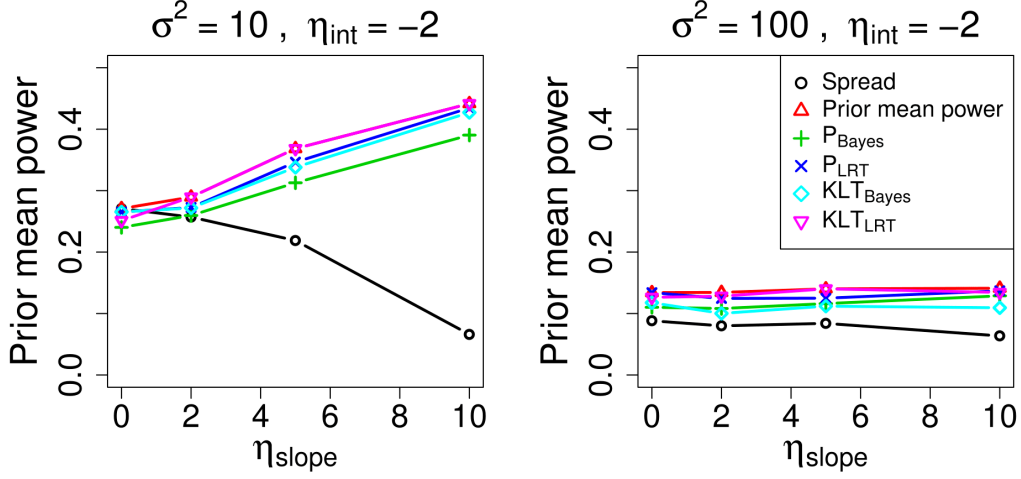


$H_1 : \theta = 1$ , where  $g_0(p) = \log(-\log(1-p))$  and  $g_1(p) = \Phi^{-1}(p)$  are the familiar complementary log-log and Probit link functions, respectively ( $\Phi$  is the standard Normal cumulative distribution function). In this model selection scenario, the coefficients  $\beta_0 = \{\beta_{0,\text{int}}, \beta_{0,\text{slope}}\}$  and  $\beta_1 = \{\beta_{1,\text{int}}, \beta_{1,\text{slope}}\}$  are nuisance parameters, and we assign the prior distribution

$$\beta_i | H_i \sim N(\eta, R), \quad (1.26)$$

for  $i = 0, 1$ . Our independent prior for  $\theta$  is Bernoulli(1/2). The design problem is to choose the design which will provide the most information for distinguishing between the two link functions.

We consider designs of 5 unique points in  $[-1, 1]$  with 100 replications of each. Within this class, we optimize the expected test information under the posterior-prior ratio and log evidence functions, i.e., (1.17) and  $\mathcal{V}(z) = \log(z)$ . Since  $\beta_0$  and  $\beta_1$  are nuisance parameters, we use the two measures in Section 1.2.3, i.e., the Bayes and MLE plug-in approaches. Under the posterior-prior ratio evidence function, we denote these two measures by  $\phi_{P_{\text{bayes}}}(M)$  and  $\phi_{P_{\text{LRT}}}(M)$ , respectively. Similarly, under the log evidence function the two measures are denoted  $\phi_{KLT_{\text{bayes}}}(M)$  and  $\phi_{KLT_{\text{LRT}}}(M)$ , respectively. The  $\phi$  notation indicates a design criterion,  $P$  indicates a connection to the expected posterior probability of  $H_1$ , (given by  $\pi_1 + \pi_0 \phi_{P_{\text{bayes}}}(M)$ ), and  $KLT$  indicates the KL divergence between the marginal data models and the testing context. The criteria are computed using Monte Carlo simulation and the optimal design under each criterion is found using a single point exchange algorithm similar to that introduced



**Figure 1.2:** Prior mean power of the likelihood ratio test under  $M_C$ , for  $C \in \{\text{Spread}, \text{Power}, P_{\text{Bayes}}, P_{\text{LRT}}, KLT_{\text{Bayes}}, KLT_{\text{LRT}}\}$ , for different settings of the priors in (1.26). For the P-optimal designs we set  $\pi_0 = \pi_1 = 0.5$ .

by [Fedorov \(1972\)](#). The design matrix optimizing criterion  $C$  is denoted  $M_C$ , for  $C \in \{P_{\text{Bayes}}, P_{\text{LRT}}, KLT_{\text{Bayes}}, KLT_{\text{LRT}}\}$ , collectively called the P-optimal and KLT-optimal designs.

We need a separate criterion by which we can evaluate and compare the optimal designs. Since power is a common quantity of interest, we choose the criterion to be the prior mean power of the likelihood ratio test, i.e.,

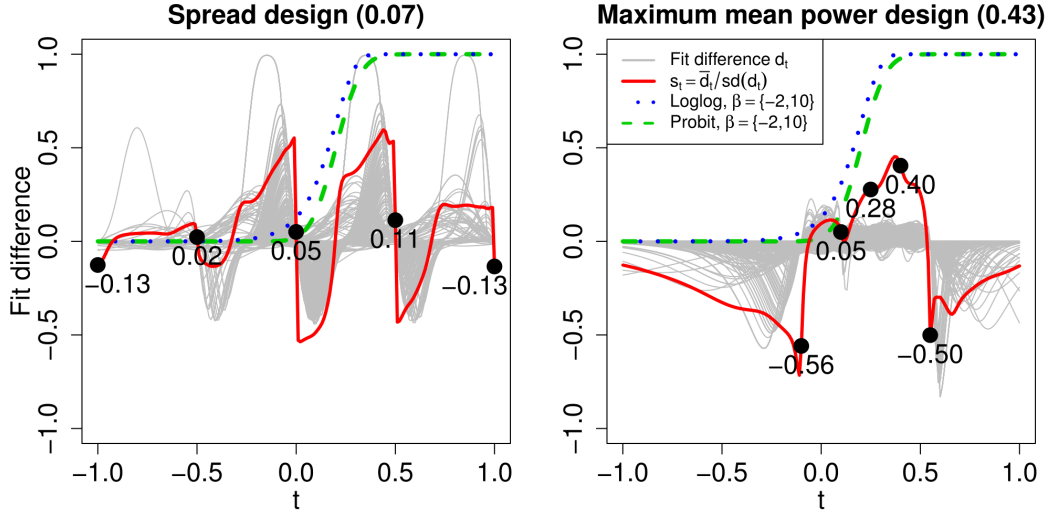
$$\int_{\Omega} \mathcal{P}(M; \theta_0, \theta_1, \beta_0, \beta_1) \psi_0(\beta_0 | \theta_0) \psi_1(\beta_1 | \theta_1) \pi_0(\theta_0) \pi(\theta_1) d(\beta_0, \beta_1, \theta_0, \theta_1), \quad (1.27)$$

where  $\Omega = \Theta_0 \times \Theta_1 \times B_0 \times B_1$ , and  $\mathcal{P}(M; \theta_0, \theta_1, \beta_0, \beta_1)$  denotes the power of the likelihood ratio test under design matrix  $M$  and given the parameters  $\theta_0, \theta_1, \beta_0, \beta_1$  (for a test size of 5%). Section 1.5.2 discusses the reasons why (1.27) or similar summaries of power are not the only measures of expected test information, or

even particularly good measures. Nonetheless, the relative familiarity of (1.27) makes it suitable for our current purpose of comparing the performance of the different optimal designs. Figure 1.2 shows the prior mean power under  $M_C$ , for  $C \in \{P_{\text{bayes}}, P_{\text{LRT}}, KLT_{\text{bayes}}, KLT_{\text{LRT}}\}$ , given various specifications of the priors in (1.26). In all cases  $R = \sigma^2 I_2$  and only  $\sigma^2$  and  $\eta$  are indicated. Note that, we tried several values of  $\eta_{\text{int}}$  but the results were qualitatively very similar, so Figure 1.2 only shows results for  $\eta_{\text{int}} = -2$ .

Also shown is the prior mean power under  $M_{\text{power}}$  and  $M_{\text{spread}}$ , the maximum prior mean power design and the spread of points  $-1, -0.5, 0, 0.5, 1$  (replicated 100 times), respectively. The P-optimal and KLT-optimal designs all perform well in terms of prior mean power, and in some cases yield considerably greater prior mean power than  $M_{\text{spread}}$ . For example, when  $\sigma^2 = 10$  and  $\eta = \{-2, 10\}$  (left panel of Figure 1.2), the design  $M_{\text{spread}}$  has prior mean power 0.07 while the P-optimal and KLT-optimal designs are all relatively close to achieving the maximum prior mean power of 0.44. The problem with  $M_{\text{spread}}$  in this case is that both inverse link functions go from 0 to 1 over a small range of the covariate  $t$  and therefore spreading the design points over the whole interval  $[-1, 1]$  is not an effective strategy.

To investigate the designs further, Figure 1.3 compares  $M_{\text{spread}}$  and  $M_{\text{power}}$ . In each plot, the design is given by the x-coordinates of the large dots (100 binary observations are recorded at each). In this illustration, 500 datasets were simulated under each design with  $\eta = \{-2, 10\}$  and  $\sigma^2 = 10$ . For reference, the complementary log-log (dotted blue line) and Probit (dashed green line) inverse



**Figure 1.3:** Comparison of  $M_{\text{spread}}$  (left) and  $M_{\text{power}}$  (right), with the parameters of (1.26) set to  $\eta = \{-2, 10\}$  and  $R = 10I_2$ . Thin grey lines show  $d_t$ , for each simulated dataset, and the thick red line shows  $s_t$  (both are described in the main text). The large dots show the design point locations (x-coordinates) and the corresponding values of  $s_t$  (y-coordinates and numbers below).

link functions are plotted for  $\beta_0 = \beta_1 = \{-2, 10\}$ . However, since  $\sigma^2 > 0$ , the actual value of  $\beta_0$  and  $\beta_1$  vary across the simulated datasets. Furthermore, for any given dataset, there is uncertainty associated with the MLE of  $\beta_0$  and  $\beta_1$ . These two sources of variation are captured by the spread of the solid thin grey lines in Figure 1.3; each corresponds to a single simulated dataset and traces the fit difference  $d_t(x^{(j)}) = g_1^{-1}(\hat{\beta}_{1,\text{int}}^{(j)} + \hat{\beta}_{1,\text{slope}}^{(j)}t) - g_0^{-1}(\hat{\beta}_{0,\text{int}}^{(j)} + \hat{\beta}_{0,\text{slope}}^{(j)}t)$  for  $t \in [-1, 1]$ , where  $\hat{\beta}_{i,\text{int}}^{(j)}, \hat{\beta}_{i,\text{slope}}^{(j)}$  are the MLEs of  $\beta_{i,\text{int}}, \beta_{i,\text{slope}}$  for dataset  $x^{(j)}$ , for  $i = 0, 1$ , and  $j = 1, \dots, 500$ . The distribution of the fit differences at any point  $t$  indicates our ability to distinguish the two inverse link functions at that point based on maximum likelihood fits. The solid thick red line summarizes by tracing the relative mean fit difference  $s_t = \bar{d}_t/\text{sd}(d_t)$ , where  $\bar{d}_t$  and  $\text{sd}(d_t)$  are the mean and standard

deviation of  $d_t$  over 500 simulations, respectively.

The y-coordinates of the large dots give the values of  $s_t$  at the design points (as do the numbers below the large dots). As expected,  $|d_t|$  is generally small at the design points, but for  $M_{\text{power}}$  the variability in  $d_t$  is low and thus  $|s_t|$  is larger at the design points than under  $M_{\text{spread}}$ . The low variability is achieved by grouping the design points together near the important steep section of the reference inverse link functions. The complementary log-log and Probit regression models fit by maximum likelihood are known to differ principally in the tails, and hence  $s_t$  is not largest at the design points in the central steep section. But, these points constrain the fits, thus reducing variability in  $d_t$  so that the two design points in the tails have large values of  $s_t$ . The designs  $M_{KLT_{\text{bayes}}}$  and  $M_{KLT_{\text{LRT}}}$  are almost identical to  $M_{\text{power}}$ , which is to be expected because intuitively the prior mean power should increase as the expected negative log Bayes factor (or likelihood ratio) increases.

### 1.2.5 Normal linear regression coefficient tests

We now discuss the Normal linear regression model

$$X|\beta, M \sim N(M\beta, \sigma^2 I), \quad (1.28)$$

and the hypotheses  $H_0 : \beta = \beta_0$  and  $H_1 : \beta \sim N(\eta, \sigma^2 R)$ . The goal is to test the adequacy of a specific value  $\beta_0$  of the regression coefficients, rather than treating them as nuisance parameters as we did in Section 1.2.4. Thus,  $\beta$  is now playing

the role of  $\theta$  in the expected test information of Definition 1.3, and formally we restrict its support under  $H_1$  to be  $B_1 = \mathbb{R}^2 / \{\beta_0\}$ . We again consider the criteria  $\phi_{P_{\text{bayes}}}(M)$  and  $\phi_{KLT_{\text{bayes}}}(M)$  which, since there are no nuisance parameters, are now simply given by Definition 1.3 under the posterior-prior ratio and log evidence function, respectively.

In the linear regression setting,  $\phi_{KLT_{\text{bayes}}}(M)$  has the closed form

$$\begin{aligned} \phi_{KLT_{\text{bayes}}}(M; \beta_0, \eta, R) &= \int_{\mathcal{X}} \log \left( \frac{f(x|H_1, M)}{f(x|\beta_0, M)} \right) f(x|H_1, M) dx \\ &= \frac{1}{2} \left( \frac{1}{\sigma^2} (\eta - \beta_0)^T M^T M (\eta - \beta_0) + \text{tr}(M^T M R) - \log(|I + M R M^T|) \right). \end{aligned} \quad (1.29)$$

The first term of (1.29) confirms our intuition that the expected test information is large when  $\beta_0$  and the mean of the alternative are well separated (with respect to  $M$ ). Heuristically, the second term of (1.29) tells us to maximize the “ratios” of the prior (alternative) variance of each parameter to the regression estimate variance. This is intuitive because we need the estimation variance to be small in comparison to the prior variance in order to effectively distinguish  $\beta$  and  $\eta$  (and hence further distinguish  $\beta$  from  $\beta_0$ ). The final term penalizes the alternative for introducing uncertainty in  $\beta$ , i.e., for avoiding exclusion of the true model by being vague.

The KLT-optimality criterion (1.29) is closely related to the D-optimality criterion popular in estimation problems (e.g., see the review by [Chaloner and Verdinelli \(1995\)](#)). The D-optimality criterion is derived from the expected es-

tion information suggested by Lindley (1956), and is given by

$$\phi_D(M) = -\log |V|, \quad (1.30)$$

where  $V = \sigma^2(M^T M + R^{-1})^{-1}$  is the posterior covariance matrix of  $\beta$  (for any value of  $X$ ). The criteria  $\phi_D$  and  $\phi_{KLT_{\text{bayes}}}$  are both entropy based, but the dependence of  $\phi_{KLT_{\text{bayes}}}$  on  $\beta_0$  and  $\eta$  distinguishes this criterion from  $\phi_D$  and other estimation focused criteria.

To gain some intuition let us consider a simple linear regression, with

$$M^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \end{pmatrix}, \quad R = \begin{pmatrix} \sigma_{\text{int}}^2 & \sigma_{\text{is}} \\ \sigma_{\text{is}} & \sigma_{\text{slope}}^2 \end{pmatrix}. \quad (1.31)$$

It is well known that in this scenario, the D-optimality criterion leads to half of the design points  $t_i$ ,  $i = 1, \dots, n$ , being at 1 and the other half at  $-1$  (or, if  $n$  is odd,  $(n+1)/2$  points at one boundary and  $(n-1)/2$  at the other). Let  $\Delta = (\eta_{\text{int}} - \beta_{0,\text{int}})(\eta_{\text{slope}} - \beta_{0,\text{slope}}) + \sigma_{\text{is}}$ , where  $\eta_{\text{int}}$  and  $\eta_{\text{slope}}$  are the mean intercept and mean slope of the alternative model, respectively. If  $\sigma_{\text{is}} = 0$ , then the sign of  $\Delta$  tells us if the lines  $\beta_{0,\text{int}} + \beta_{0,\text{slope}}t$  and  $\eta_{\text{int}} + \eta_{\text{slope}}t$  have greater separation at  $-1$  or at 1. For any  $\sigma_{\text{is}}$ , it is easily shown that  $\phi_{KLT_{\text{bayes}}}$  is optimized by placing all points at 1 if  $\Delta > 0$ , by placing them at  $-1$  if  $\Delta < 0$ , and by any design dividing the points between the boundaries if  $\Delta = 0$ . Generally, designs based on test information measures trade robustness for power in distinguishing particular models, and the behavior just described is an instance of the inevitable sensitivity to the hypotheses

mentioned in Section 1.1.1. However, in the current context, we found that designs optimizing  $\phi_{P_{\text{bayes}}}$  are slightly more robust than those optimizing  $\phi_{KLT_{\text{bayes}}}$  in that they divide the points between both the boundaries, unless the hypotheses are far more separated at one boundary than at the other.

## 1.3 Observed test information in theory and application

### 1.3.1 Observed test information: building blocks

Observed test information is key in practice when we have observed some data and want to know how much information they contain in order to decide if we should collect more. It is also important conceptually because it is the implicit building block for expected and conditional test information.

First consider the estimation information introduced by DeGroot (1962) and reviewed in Section 1.1.3. After an experiment  $\xi$  is conducted, the observed estimation information gained is the reduction in uncertainty,  $U(\pi) - U(p(\cdot|x))$ , where  $x \in \mathcal{X}$  is the observed outcome. Observed estimation information is not necessarily non-negative because, by chance, after observing  $x$  we may have more uncertainty about  $\theta$  as measured by  $U$ , e.g., the posterior may be more diffuse than the prior due to likelihood-prior conflict; see Reimherr et al. (2014). (This posterior “dilation” can even be deterministic; see Seidenfeld and Wasserman (1993).) Interestingly, DeGroot (1962) did not explicitly mention observed information,



but Lindley (1956) did define it (as above) in the case where  $U$  is the entropy function (1.1). Ginebra (2007) restricted all observed information measures to be positive and interprets them to be capturing model checking information, in addition to information about  $\theta$ , but does not explain why non-negativity can ensure this interpretation is reasonable. From a Bayesian perspective, the definition given by Lindley (1956) is valid, and we therefore take this as the basis of observed estimation information.

Following an analogous approach to Lindley (1956), we define observed test information in Definition 1.5 (below) by simply removing the expectation appearing in the expected test information of Definition 1.3. However, the resulting relationship between observed and expected information is more subtle than in the estimation case. Indeed, Definition 1.3 conditions on  $H_1$  to average over the unobserved data, but the actual data used in Definition 1.5 may be generated under  $H_0$ .

**Definition 1.5** *The observed test information provided by the experiment  $\xi$  for comparing the hypotheses  $H_0 : \theta \in \Theta_0$  and  $H_1 : \theta \in \Theta_1$ , for a given evidence function  $\mathcal{V}$  and a proper prior  $\pi$ , is defined as*

$$\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi, x) = \mathcal{V}(1) - \mathcal{V}(\text{BF}(x|H_0, H_1)), \quad (1.32)$$

where  $x$  is the observed outcome of  $\xi$ , and  $\Theta_0 \cap \Theta_1 = \emptyset$ .

Since Bayesians condition on observed data, the prefix ‘observed’ is redundant, but it is retained for clarity. The quantity defined by (1.32) is not necessarily

non-negative. However, it is positive when  $\mathcal{V}$  is increasing and the Bayes factor favors  $H_1$ , i.e.,  $\text{BF}(x|H_0, H_1) < 1$ . Often, it seems sensible for  $\mathcal{V}$  to be increasing because we want (1.32) to increase as the Bayes factor decreases towards zero (since observed test information should be compatible with Definition 1.3 which assumes  $H_1$ ). For  $\mathcal{V}$  increasing, a negative value of  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi, x)$  indicates that the evidence in the observed data supports  $H_0$ , either because  $H_0$  is in fact the more accurate hypothesis or due to chance. Since the data can only support one of the hypotheses, for increasing  $\mathcal{V}$  it follows that exactly one of the dual observed test information measures  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi, x)$  and  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0, \pi, x)$  will be positive (unless they are both zero). Also, usually only one of  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi, x)$  and  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0, \pi, x)$  will reasonably approximate the corresponding expected test information,  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi)$  and  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0, \pi)$ , respectively.

In the case of the sharp hypotheses  $H_0 : \theta = \theta_0$  and  $H_1 : \theta = \theta_1$ , the Bayes factor in (1.32) becomes the usual likelihood ratio  $\text{LR}(\theta_0, \theta_1|x)$ , and plugging in the MLE of  $\theta$  for  $\theta_1$  ensures the measure is non-negative, again provided  $\mathcal{V}$  is increasing. A more complete frequentist approach than plugging in the MLE of  $\theta$  is to provide a confidence interval for  $\mathcal{V}(1) - \mathcal{V}(\text{LR}(\theta_0, \theta|x))$  based on a confidence interval for  $\theta$ .

We highlight that, in the current observed data case, our use of dual measures is again key because it ensures a symmetric treatment of the hypotheses, which is not easily achieved by other means. For example, consider a  $\mathcal{V}$  that is concave, increasing, and passes through  $\{0, 1\}$  (for all non-zero prior probabilities  $\pi_0$  and  $\pi_1$  whose sum is one), then  $\mathcal{V}(z) + z\mathcal{V}(1/z)$  is also concave, gives expected test information

$\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi) + \mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0, \pi)$ , and yields non-negative observed test information, as required by [Ginebra \(2007\)](#). However, in many cases, excluding the case of (1.20), this approach does not treat the hypotheses equally. For example, we can modify the evidence function  $\mathcal{V}(z) = \sqrt{z} - 1$  to  $\mathcal{V}(z) + z\mathcal{V}(1/z) = 2\sqrt{z} - 1 - z$ , but the resulting observed test information has a maximum of one for data supporting  $H_1$ , and is unbounded for data supporting  $H_0$ . Our approach using dual observed test information measures is therefore more appealing because  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_0, H_1, \pi, x)$  and  $\mathcal{I}_{\mathcal{V}}^T(\xi; H_1, H_0, \pi, x)$  are symmetrically defined. This symmetry is even more foundational than the coherence identity (1.15) because observed test information identifies the underlying statistics of interest.

Next, in the same spirit, we define the observed conditional test information provided by conducting the experiment  $\xi_2$  after observing the outcome  $x_1$  of an experiment  $\xi_1$  to be

$$\mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1; H_0, H_1, \pi, x_1) = W(x_1) - E_{X_2}[W(x_1, X_2)|H_1, x_1]. \quad (1.33)$$

This is simply the information given in Definition 1.4, but without an expectation over  $x_1$ , because it has been observed. In sequential design we require a version of the coherence identity (1.16) to hold for (1.33). In particular, given some observed data  $x_1$ , we do not want the optimality of our design for new data to depend on which hypothesis is true. If we assume the symmetry condition (1.15), then it straightforwardly follows that

$$\mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1; H_0, H_1, \pi, x_1) = z(x_1)\mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1; H_1, H_0, \pi, x_1), \quad (1.34)$$

for all  $x_1 \in \mathcal{X}$ , where  $z(x_1) = \text{BF}(x_1|H_0, H_1)$ . The factor  $z(x_1)$  appears in (1.34) (but not in (1.15)) because the observed data  $x_1$  already favors one of the hypotheses before any new data are collected. If (1.34) holds, then the design that optimizes  $\mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1; H_0, H_1, \pi, x_1)$  also optimizes  $\mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1; H_1, H_0, \pi, x_1)$ , meaning that, as required, we do not need to know which hypothesis is true in order to find a good choice of  $\xi_2$ .

### 1.3.2 Sequential design for linear regression coefficient tests

Consider the linear regression model (1.28) introduced in Section 1.2.5 and the test of  $H_0 : \beta = \beta_0$  against  $H_1 : \beta \sim N(\eta, R)$  (i.e.,  $\sigma^2 = 1$ ). Given some initial observed data  $x_{\text{ob}}$ , the sequential design problem is to choose a design matrix  $M_{\text{mis}}$  for additional data  $X_{\text{mis}}$ . In our simulation study, we generate a parameter vector  $\beta$  under  $H_1$  and then simulate the initial observed data  $x_{\text{ob}}$  according to a cubic regression model of the form (1.28), i.e., the design matrix is

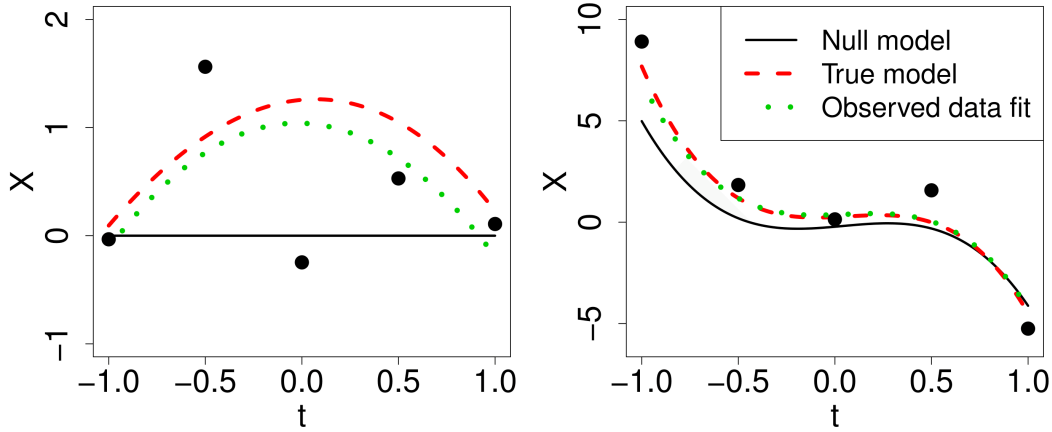
$$M_{\text{ob}}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_{n_{\text{ob}}} \\ t_1^2 & t_2^2 & \cdots & t_{n_{\text{ob}}}^2 \\ t_1^3 & t_2^3 & \cdots & t_{n_{\text{ob}}}^3 \end{pmatrix}, \quad (1.35)$$

where  $t_i \in [-1, 1]$  are the design points, for  $i = 1, \dots, n_{\text{ob}}$ . Specifically, we set  $n_{\text{ob}} = 5$  and the observed data design points  $t_i$ ,  $i = 1, \dots, 5$ , are  $-1, -0.5, 0, 0.5, 1$ . Examples of  $x_{\text{ob}}$  are plotted in Figure 1.4.

Given the observed data,  $n_{\text{mis}} = 5$  new design points are chosen by optimizing

the observed conditional test information (1.33) with respect to the design matrix  $M_{\text{mis}}$ , under the posterior-prior ratio and log evidence functions. That is, we optimize the conditional versions of the P-optimality and KLT-optimality criteria discussed in Section 1.2.5. The conditional P-optimality criterion is the expected reduction in posterior probability of the null when we collect  $X_{\text{mis}}$  relative to its prior probability. We approximate it using a Monte Carlo estimate, under the prior probabilities  $\pi_0 = \pi_1 = 0.5$ . The conditional KLT-optimality criterion is straightforwardly given by  $\log(z(x_{\text{ob}})) + \phi_{KLT_{\text{bayes}}}(M; \beta_0, \eta_{\text{ob}}, V_{\text{ob}})$ , where  $\phi_{KLT_{\text{bayes}}}$  is specified in (1.29), and  $\eta_{\text{ob}}$  and  $V_{\text{ob}}$  are the observed data posterior mean and covariance matrix of  $\beta$  under  $H_1$ , respectively. For comparison, we also optimize the conditional D-optimality criterion  $\log |V_{\text{ob}}| + \log \left| (M_{\text{mis}})^T M_{\text{mis}} + V_{\text{ob}}^{-1} \right|$  with respect to  $M_{\text{mis}}$ . Often it is not clear how to use the D-optimality criterion and other estimation based criteria to choose designs for testing, but the current scenario is an exception because the hypotheses are nested.

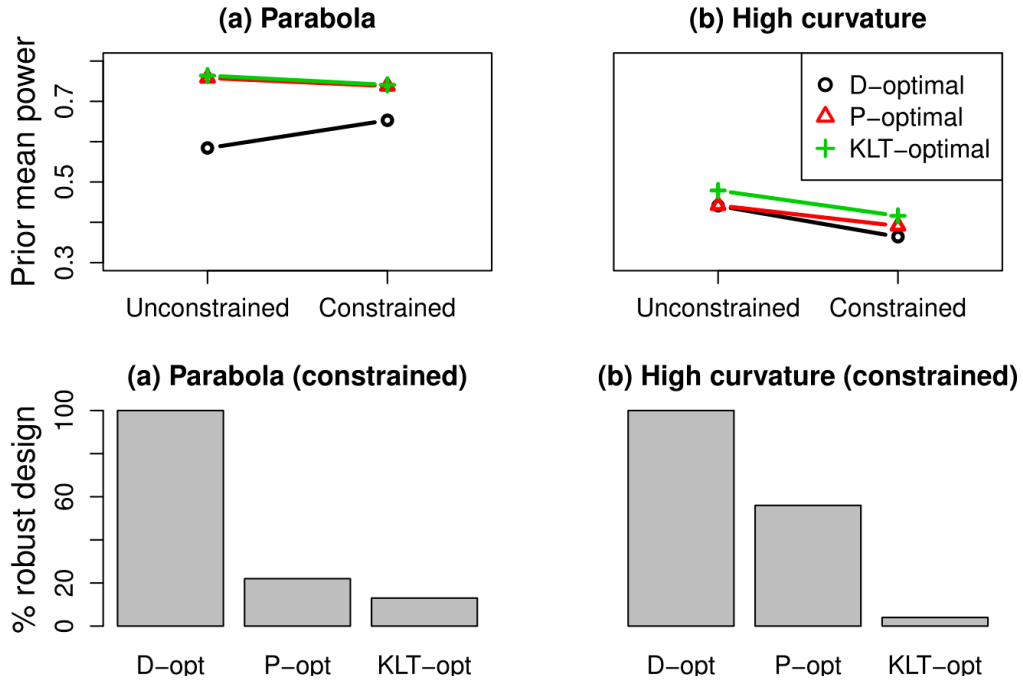
To generalize beyond a single value of  $\beta$ , we generate  $\beta^{(j)} \sim N(\eta, R)$ , for  $j = 1, \dots, 100$ , and for each  $j$  we generate observed datasets  $x_{\text{ob}}^{(j,k)}$ , for  $k = 1, \dots, 250$ . Then, for each simulated dataset  $x_{\text{ob}}^{(j,k)}$ , we find the conditional P-optimal, KLT-optimal, and D-optimal design for the missing data  $X_{\text{mis}}$ . To compare performance, we also calculate the prior mean power (1.27) of the likelihood ratio test under each of these three procedures. In our simulations, we set  $R = 0.2I_4$ , and use various values of  $\eta$  and  $\beta_0 = (\beta_{0,\text{int}}, \beta_{0,\text{lin}}, \beta_{0,\text{quad}}, \beta_{0,\text{cubic}})$ . First, Figure 1.5 part (a) corresponds to simulations with  $\beta_0 = (0, 0, 0, 0)$  and  $\eta = (1.1, 0, -1.3, 0)$  (i.e., the alternative mean model is parabola shaped). For these choices, the maximum



**Figure 1.4:** The null and true cubic regression models and the observed data posterior mean fit. The observed data are indicated by large dots. The left and right plots show example simulations used in producing parts (a) and (b) of Figure 1.5, respectively.

separation between the null and true model is usually not at the boundaries of the interval  $[-1, 1]$ ; see the example simulation given on the left of Figure 1.4. The top row of Figure 1.5 shows the prior mean power of the three procedures when any design points in  $[-1, 1]$  are allowed (i.e.,  $M_{\text{mis}}$  is unconstrained) and also when only two possibilities for  $M_{\text{mis}}$  are allowed (these latter results are for the constrained optimization example discussed shortly). For part (a), the conditional D-optimal procedure performs relatively poorly because it almost invariably places all the new points at the boundaries, a good strategy for estimation but not for hypothesis testing. The conditional P-optimality and KLT-optimality procedures instead place the points near  $t = 0$ , and consequently are substantially superior in terms of prior mean power.

For the simulations corresponding to Figure 1.5 part (b), we first generated the null parameters  $\beta_{0,\text{int}}^{(j)}, \beta_{0,\text{lin}}^{(j)} \sim \text{Uniform}(-1, 1)$  and  $\beta_{0,\text{quad}}^{(j)}, \beta_{0,\text{cubic}}^{(j)} \sim \text{Uniform}(-10, 10)$ ,



**Figure 1.5:** Prior mean power of the likelihood ratio test under the conditional D-optimality, P-optimality, and KLT-optimality procedures, across 250 datasets simulated under  $\beta \sim N(\eta, 0.2I_4)$  (first row, unconstrained values). The main text describes the generation of  $\beta_0$  and  $\eta$  for parts (a) and (b). The first row constrained values show the prior mean powers when the only missing data designs allowed are (i) and (ii) (see the main text). For this case, the second row shows the percentage of simulations in which design (i) was selected.

and then we set  $\eta^{(j)} = \beta_0^{(j)}$  and drew  $\beta^{(j)} \sim N(\eta^{(j)}, R)$ , for  $j = 1 \dots 100$ . Under these settings, the maximum separation between the null curve and the observed data posterior mean fit tends to be at one of the boundaries. Hence, the conditional D-optimality procedure performs reasonably, because it again divides the points between the two boundaries. Thus, in part (b) of Figure 1.5 the three procedures perform similarly.

We now briefly investigate how the three criteria perform if we impose some robustness to model misspecification. The points labeled “constrained” in the

first row of Figure 1.5 show the prior mean power of the likelihood ratio test when the three criteria are used to choose between two missing data designs: (i) the spread of points  $\mathbf{t}_{\text{spread}} = \{-1, -0.5, 0, 0.5, 1\}$ ; (ii) the narrower spread of points  $\frac{1}{5}\mathbf{t}_{\text{spread}} + \text{sep}_{\text{max}}$ , where  $\text{sep}_{\text{max}}$  is the location of maximum separation between the observed data posterior mean model and the null mode. If  $\text{sep}_{\text{max}}$  is near a boundary then all the points are shifted left or right to avoid any crossing the boundary, but they still cover an interval of length 0.4. The results follow a similar pattern to before, except that now the prior mean power is usually lower, principally because designs placing all the points at a single location have been excluded. The first row of Figure 1.5 shows that the constrained conditional P-optimality procedure has prior mean power almost as high as the constrained conditional KLT-optimality procedure, but the second row indicates that it also selects the more robust design (i) far more often (usually when the posterior probability of  $H_1$  is low). Thus, the conditional P-optimality procedure offers a compromise between power for distinguishing the hypotheses of interest and robustness.

## 1.4 Links between test and estimation information

### 1.4.1 Fraction of observed test information

Nicolae et al. (2008) propose several measures of the fraction of observed test



information to guide data collection decisions in genetic linkage studies (see Section 1.1.2). In words, the fraction of observed test information is the ratio of the observed test information and an estimate of the test information based on the complete data. We provide the general mathematical form in Definition 1.6 (below) because the fraction of observed test information is important in sequential design and for establishing theoretical connections between test and estimation information. In sequential design, it gives the relative amount of test information still obtainable (under the alternative), and can also help in assessing the difficulty of collecting each unit of that remaining test information. Analogously, finding a specific subgroup of a population is on average more difficult per unit if the subgroup is small compared to the population.

**Definition 1.6** *The fraction of observed test information provided by the first part of the composite experiment  $\xi = (\xi_1, \xi_2)$  for comparing the hypotheses  $H_0 : \theta \in \Theta_0$  and  $H_1 : \theta \in \Theta_1$ , for a given evidence function  $\mathcal{V}$  and a proper prior  $\pi$ , is defined as*

$$\mathcal{FI}_{\mathcal{V}}^T(\xi_2|\xi_1; H_0, H_1, \pi, x_1) = \frac{\mathcal{I}_{\mathcal{V}}^T(\xi_1; x_1)}{\mathcal{I}_{\mathcal{V}}^T(\xi_1; x_1) + \mathcal{I}_{\mathcal{V}}^T(\xi_2|\xi_1; x_1)}, \quad (1.36)$$

where  $x_1$  is the observed outcome of  $\xi_1$ , and  $\Theta_0 \cap \Theta_1 = \emptyset$ .

If  $\mathcal{I}_{\mathcal{V}}^T(\xi_1; x_1) \geq 0$  then it follows that (1.36) is between 0 and 1. In practice, if  $\mathcal{FI}_{\mathcal{V}}^T(\xi; x_1)$  is close to one then we may decide not to perform  $\xi_2$ , particularly if it is expensive. The canonical example sets  $\mathcal{V}(z) = \log(z)$  and thus takes the ratio of the observed data log Bayes factor and the expected complete data log Bayes

factor. Similarly, in the frequentist case, [Nicolae et al. \(2008\)](#) suggest the measure

$$\mathcal{R}I_1 = \frac{\log \text{LR}(\theta_{\text{ob}}, \theta_0 | x_{\text{ob}})}{E_{X_{\text{mis}}}[\log \text{LR}(\theta_{\text{ob}}, \theta_0 | X_{\text{ob}}, X_{\text{mis}}) | \theta_{\text{ob}}, x_{\text{ob}}]}, \quad (1.37)$$

where  $\theta_{\text{ob}}$  is the MLE of  $\theta$  based on  $x_{\text{ob}}$ .

The decision whether to collect more data depends on which hypothesis is true, because if the observed data supports the false hypothesis then our need for additional data is greater. Thus, it is unsurprising that there is no general coherence identity for the fraction of observed test information. In practice, we suggest using  $\mathcal{F}\mathcal{I}_{\mathcal{V}}^T(\xi_2 | \xi_1; H_0, H_1, \pi, x_1)$  if  $z(x_1) \leq 1$  and  $\mathcal{F}\mathcal{I}_{\mathcal{V}}^T(\xi_2 | \xi_1; H_1, H_0, \pi, x_1)$  otherwise. The resulting measure has a similar interpretation as (1.36) but takes account of which hypothesis is more likely, and is always between 0 and 1. In the special case where  $\mathcal{V}(1) = 0$  (and (1.15) is satisfied), we have  $\mathcal{F}\mathcal{I}_{\mathcal{V}}^T(\xi_2 | \xi_1; H_0, H_1, \pi, x_1) = \mathcal{F}\mathcal{I}_{\mathcal{V}}^T(\xi_2 | \xi_1; H_1, H_0, \pi, x_1)$ , but this is not a coherence identity since the corresponding observed test information is negative on one side of the equality and positive on the other.

## 1.4.2 Connections between estimation and test information

[Meng and van Dyk \(1996\)](#) show that the *relative augmentation function*,

$$\mathcal{R}I(\theta) = \frac{\log \text{LR}(\theta_{\text{ob}}, \theta | x_{\text{ob}})}{E_{X_{\text{co}}}[\log \text{LR}(\theta_{\text{ob}}, \theta | X_{\text{co}}) | \theta_{\text{ob}}, x_{\text{ob}}]}, \quad (1.38)$$

converges to the fraction of observed Fisher information,

$$\mathcal{R}I_E = \frac{I_{\text{ob}}}{I_{\text{ob}} + I_{\text{mis}}}, \quad (1.39)$$

as  $|\theta - \theta_{\text{ob}}| \rightarrow 0$ . Here,  $I_{\text{ob}}$  is the usual observed Fisher information, and  $I_{\text{mis}}$  is the missing Fisher information which is given by

$$I_{\text{mis}} = E_{X_{\text{co}}} \left[ - \frac{\partial^2 \log f(X_{\text{co}} | x_{\text{ob}}, \theta)}{\partial \theta^2} \Big| x_{\text{ob}}, \theta \right] \Big|_{\theta = \theta_{\text{ob}}}. \quad (1.40)$$

As [Nicolae et al. \(2008\)](#) point out, replacing  $\theta$  with  $\theta_0$  gives us the same limit for the measure  $\mathcal{R}I_1$  in (1.37). This result is intuitive in that we might expect test information to coincide with estimation information when the two hypotheses are both very close to  $\theta_{\text{ob}}$ . The following theorem generalizes the equivalence, and its proof is given in Appendix A.1.

**Theorem 1.1** *Let the hypotheses be  $H_0 : \theta = \theta_0$  and  $H_1 : \theta = \theta_1$ , and suppose that the derivatives of the evidence function  $\mathcal{V}$  exist at 1. Then, for univariate  $\theta$  and  $\theta_1 = \theta_{\text{ob}}$ , we have*

$$\mathcal{F}\mathcal{I}_{\mathcal{V}}^T(\xi_2 | \xi_1; H_0, H_1, x_{\text{ob}}) = \frac{\mathcal{V}'(1)I_{\text{ob}}}{\mathcal{V}'(1)I_{\text{ob}} - \mathcal{V}''(1)I_{\text{mis}}} + O_p(|\theta_0 - \theta_{\text{ob}}|), \quad (1.41)$$

*under the uniform integrability condition given in the proof in Appendix A.1.*

It is possible to extend Theorem 1.1 to avoid the univariate condition and sharp hypotheses (by using priors that converge to delta functions), but the current form suffices to illustrate the connection between test and Fisher information.

The theorem tells us that if  $\mathcal{V}'(1) = -\mathcal{V}''(1)$ , then  $\mathcal{FT}_{\mathcal{V}}^T(\xi_2|\xi_1; H_0, H_1, x_{\text{ob}})$  will exactly correspond to  $\mathcal{RI}_E$  as  $|\theta_0 - \theta_{\text{ob}}| \rightarrow 0$ . Otherwise, the *relative conversion number*  $C_{\mathcal{V}} = -\mathcal{V}''(1)/\mathcal{V}'(1)$  indicates how much of the missing data estimation information is converted to test information in the limit, relative to the conversion of observed estimation information. For example, under the posterior-prior ratio evidence function (1.17) we have  $C_{\mathcal{V}} = 2\pi_0$ , and therefore the stronger our initial bias in favor of the null, the greater the importance of the missing data estimation information, relative to the observed estimation information. This makes sense because Fisher information measures our ability to estimate the true parameter, and the value of successful estimation for testing depends on the strength of our prior separation of the hypotheses. If  $\pi_0 = 0.5$ , then all estimation information will be helpful because the prior does not separate the hypotheses, hence  $C_{\mathcal{V}} = 1$  and the fraction of observed test and estimation information coincide. When  $\pi_0$  is close to 0, the posterior probability of  $H_1$  (the hypothesis assumed true by Definition 1.6) will be close to one, even though the observed data provides no evidence. Thus, we have  $C_{\mathcal{V}} \approx 0$ , because there is little to be gained by collecting more data. When  $\pi_0$  is close to 1, the prior is in conflict with our assumption that  $H_1$  is true, and therefore estimation information from new data not only has the potential to distinguish the hypotheses, but also to overcome false information from the prior. Hence, we expect  $C_{\mathcal{V}} > 1$ . In the current example, it turns out that  $C_{\mathcal{V}} \approx 2$  because our choice of  $\mathcal{V}$  sets  $C_{\mathcal{V}} \propto \pi_0$  and  $C_{\mathcal{V}} = 1$  for  $\pi_0 = 0.5$ .

The relative conversion number has similar interpretations for other evidence functions. In each case it indicates the relative worth of the missing data esti-

mation information for testing, when there is no evidence in the observed data. Thus,  $C_{\mathcal{V}}$  provides a characterization of the general approach to testing implied by the evidence function, i.e., whether we would be likely to collect additional data if the observed data did not separate the hypotheses.

## 1.5 Discussion and further work

### 1.5.1 Classification

We now briefly explain how our test information framework extends to classification problems. Consider a classification problem with  $m$  classes (hypotheses) denoted  $A_i$ , for  $i = 1, \dots, m$ , and let  $z_{ij}(X)$  be the Bayes factor  $f(X|A_i)/f(X|A_j)$ , for  $i = 1, \dots, m$ ,  $j = 2, \dots, m$ . The generalized evidence function  $\mathcal{V}_m(\mathbf{z}_{\cdot j}; A_i, A_j)$  quantifies the evidence for class  $A_i$  under class  $A_j$ , and is a function of  $\mathbf{z}_{\cdot j} = \{z_{ij} : i \in \{1, \dots, m\} \setminus \{j\}\}$  (and the class prior probabilities  $\pi_1, \dots, \pi_m$ ). Thus, under  $A_j$ , the expected test information that is due to a reduction in evidence for  $A_i$  is given by

$$\mathcal{I}_{\mathcal{V}_m}^T(\xi; A_i, A_j) = \mathcal{V}_m(\mathbf{1}_{m-1}; A_i, A_j) - E_X[\mathcal{V}_m(\mathbf{z}_{\cdot j}(X); A_i, A_j) | A_j], \quad (1.42)$$

where  $\mathbf{1}_{m-1}$  is a vector of  $m - 1$  ones. For example, the posterior-prior ratio evidence function can be generalized to

$$\mathcal{V}_m(\mathbf{z}_{\cdot j}; A_i, A_j) = \frac{z_{ij}}{\pi_j + \sum_{k \neq j} \pi_k z_{kj}}, \quad (1.43)$$

in which case (1.42) is the expected difference between the prior and posterior probability of  $A_i$ , relative to its prior probability, under the assumption that  $X$  will be generated under  $A_j$ . Since the information for separating each pair of classes is important, we define the total expected test information to be

$$\mathcal{I}_{\mathcal{V}_m}^T(\xi; A_1, \dots, A_m) = \sum_{j=2}^m \sum_{i=1}^{j-1} \mathcal{I}_{\mathcal{V}_m}^T(\xi; A_i, A_j). \quad (1.44)$$

It is straightforward to show that, for any concave generalized evidence function  $\mathcal{V}_m$ , there exists a concave function  $\mathcal{V}_m^*$  of  $\mathbf{z}_{\cdot m}$  such that  $\mathcal{V}_m^*(\mathbf{1}_{m-1}) - E_X[\mathcal{V}_m^*(\mathbf{z}_{\cdot m}(X)) | A_m]$  equals  $\mathcal{I}_{\mathcal{V}_m}^T(\xi; A_1, \dots, A_m)$ . Therefore, (1.44) is a genuine extension of our framework.

Next, if  $\mathcal{V}_m$  satisfies a natural extension of the symmetry condition (1.15), namely  $\mathcal{V}_m(\mathbf{z}_{\cdot j}; A_i, A_j) / \mathcal{V}_m(\mathbf{z}_{\cdot i}; A_j, A_i) = z_{ij}$ , then we have  $\mathcal{I}_{\mathcal{V}_m}^T(\xi; A_i, A_j) = \mathcal{I}_{\mathcal{V}_m}^T(\xi; A_j, A_i)$ , for  $i = 1, \dots, m$ , and  $j = 1, \dots, m$ , e.g., one such  $\mathcal{V}_m$  is (1.43). From this condition follows the generalized coherence identity

$$\mathcal{I}_{\mathcal{V}_m}^T(\xi; A_{k_1}, \dots, A_{k_m}) = \mathcal{I}_{\mathcal{V}_m}^T(\xi; A_{k'_1}, \dots, A_{k'_m}), \quad (1.45)$$

for all choices of  $(k_1, \dots, k_m)$  and  $(k'_1, \dots, k'_m)$  that are permutations of  $(1, \dots, m)$ . A generalized identity similar to (1.34) also holds for the observed conditional test information corresponding to (1.44). In classification problems the practical importance of such coherence identities becomes greater because there can be many classes and often there is not a single special class that we are willing to assume true. They are also again fundamental, and we cannot, for example,

forgo (1.45) and instead replace each term on the right hand side of (1.44) by  $\mathcal{I}_{\mathcal{V}_m}^T(\xi; A_i, A_j) + \mathcal{I}_{\mathcal{V}_m}^T(\xi; A_j, A_i)$ , because this leads to *observed* test information measures that treat the classes asymmetrically, see Section 1.3.1.

## 1.5.2 Discussion

In this paper, we propose a general framework for constructing test information measures and illustrate their use in experimental design. Our simulation studies show that, for linear regression coefficient tests, test information based designs perform better than D-optimal designs. In the case of non-nested model selection, it is unclear how to even construct estimation information based designs, but the approach for test information based designs is straightforward, see Section 1.2.4. Additionally, we identify an appealing posterior probability based test information measure that has an intuitive interpretation, satisfies our fundamental coherence identities, and is easily generalized for use in classification problems.

Our framework also helpfully rules out various measures of test information. For example, the variance of the log Bayes factor is essentially a measure that [Nicolae et al. \(2008\)](#) rejected after some consideration, but with our framework rejection is immediate because the evidence function corresponding to this measure is of the form  $\mathcal{V}(z) = -(\log z - c)^2$ , which is not concave. Perhaps the most notable quantity ruled out is power, and we now briefly discuss why it is omitted. The obvious problem with power is that it is unclear how to calculate it in the presence of composite hypotheses and nuisance parameters. The underlying power surface can be summarized by quantities such as the prior mean power (1.27), but these

summaries lose the frequentist interpretation, and can be particularly expensive to compute. The second major problem is that there is not an intuitive measure of *observed* power. Without a coherent relationship between observed and “expected” power, it is difficult to see how sequential design decisions (e.g., stopping rules) could reasonably be based on power. A further fundamental difficulty is that power does not have the maximal information interpretation discussed in Section 1.2.1 because it incorporates an investigator-specific critical region.

### 1.5.3 Future work

A natural direction for future work is to investigate how test and estimation information measures can be combined to find designs that are good for both testing and estimation. Some work has been done along these lines by [Borth \(1975\)](#) in the special case of the entropy approach taken by [Lindley \(1956\)](#) and [Box and Hill \(1967\)](#). However, in general, test and estimation information are not related simply, and therefore trying to directly find designs that are good for both testing and estimation may not be an effective strategy. Instead, we can divide up the design points and construct two designs, one that is good for testing and one that is good for estimation. The overall design should then have reasonable properties for both problems. Future work will be to explore this approach and investigate methods for setting the proportion of the design points to be allocated to each problem.



# 2

## Disentangling overlapping astronomical sources using spatial and spectral information

### **2.1 Introduction**

When two or more sources are situated close enough to each other that there is a substantial overlap of their Point Spread Functions (PSFs), they pose a many-fold problem to astronomical analysis. The first is to recognize that there is an overlap,

the second is to determine the number of distinct sources that are involved, the third is to measure their relative intensities, and the fourth is to separate them sufficiently to be able to carry out useful secondary analyses like spectral fitting and variability analysis. These problems are especially complicated for high-energy photon detectors, since the data are firmly in the Poisson regime, background is often a significant component of the data, and the simplifying approximations of a Gaussian process are usually inapplicable. Many researchers have considered the simpler problem of a single source contaminated by background in the low counts regime (e.g., [Kraft et al. 1991](#), [Loredo 1992](#), [van Dyk et al. 2001](#), [Park et al. 2006](#), [Weisskopf et al. 2007](#), [Laird et al. 2009](#), [Knoetig 2014](#), [Primini and Kashyap 2014](#)), and have generally found that Poisson-likelihood based Bayesian techniques are well suited to address this category of problems.

However, in the case of multiple sources, progress has been slow, and the choices limited. One could construct approximate measures of intensities of the component sources in the Gaussian regime via matrix inversion ([Kashyap et al. 1994](#)), or choose to minimize contamination by limiting the sizes of the apertures to cover only the cores of the PSFs ([Broos et al. 2010](#)), or carry out full-fledged 2-D spatial modeling. All these are approximate or computationally intensive solutions. An important advance was made recently by [Primini and Kashyap \(2014\)](#), who developed a fully Bayesian aperture photometry method that simultaneously models the intensities of the overlapping sources and the intensity of the background. Their method can be applied to any counts image with multiple overlapping sources, with a practical computational limit of up to five sources.

**Table 2.1:** Symbols used in this chapter. Notation used only in a single section is defined where it appears and is not included in this table.

Symbol	Definition
$(x_i, y_i)$	Location of photon $i$ on the detector
$E_i$	Energy (PI channel) of photon $i$
$\mu_j$	True location of source $j$ (2-D coordinates)
$f_{\mu_j}$	Point Spread Function centered at $\mu_j$
$\alpha_j$	Spectral shape parameter for source $j$ (full model)
$\gamma_j$	Spectral mean parameter for source $j$ (full model)
$\alpha_{jl}$	Spectral shape parameter $l$ for source $j$ (extended full model)
$\gamma_{jl}$	Spectral mean parameter $l$ for source $j$ (extended full model)
$\pi_{jl}$	Weight of <i>gamma</i> component $l$ in source $j$ spectral model (extended full model)
$E_{\min}, E_{\max}$	Minimum and maximum detected energy (PI channel)
$w_j$	Relative intensity of source $j$ ( $j = 0$ for background)
$K$	True number of sources ( $K_{\text{true}}$ for emphasis)
$k$	A possible value of $K$
$\kappa$	Prior mean of $K$
$s_i$	True source of photon $i$ (takes the values $0, \dots, K$ , with 0 indicating background)
$n_j$	True number of photons detected from source $j$ ( $j = 0$ for background)
$n$	Total number of photons detected i.e. $\sum_{j=0}^K n_j$
$\theta_j$	Full model parameters for source $j$ i.e. $\{w_j, \mu_j, \alpha_j, \gamma_j\}$
$\Theta_K$	All full model source specific parameters i.e. $\{\theta_0, \dots, \theta_K\}$ where $\theta_0 = w_0$
$L^{\text{full}}, L^{\text{sp}}, L^{\text{ext}}$	Likelihood function of the full, spatial-only, and extended full models, respectively
$\psi^{(t)}$	The value of generic parameter $\psi$ in iteration $t$ of the algorithm
$x, y, E, s$	Vectors of the corresponding photon specific variables (see earlier table entries)
$I_A$	Indicator function equal to 1 if the event $A$ occurs (e.g. $K = 3$ ) and 0 otherwise

Despite this, most of the problems listed above are still extant.

Typically, X-ray data are collected as lists of events, with each event tagged by its location on the detector, its energy,<sup>1</sup> and its arrival time. Binning the positions into images causes a loss of information that could be alleviated by carrying out the analysis on the unbinned event lists. In such a case, it becomes feasible to disentangle individual events and allocate them probabilistically to the several

---

<sup>1</sup>The detector records the pulse height amplitude (PHA), which is roughly proportional to the energy of the incoming photon. These values are often reported as pulse-invariant (PI) gain-corrected PHAs. The distribution of PI for a photon at a given energy is encoded in the detector’s Redistribution Matrix File (RMF). In the following, we use “energy” as a synonym for this recorded PI, and clarify only if there is any ambiguity.

sources that comprise the dataset. In the following, we describe an algorithm that directly addresses three of the four problems listed above: it dynamically determines the number of overlapping sources, measures their intensities, and pools individual events into clusters for which follow-up spectral analysis can be carried out. There are related approaches for longer wavelength data originating from an unknown number of sources, for example, [Brewer et al. 2013](#) and [Safarzadeh et al. 2014](#). The former uses Gaussian process models to identify stellar oscillation modes, and the latter uses simulated Herschel images based on Hubble data to investigate a disentangling method. The principal difference between these methods and our approach is that they conduct analysis at the pixel level, whereas we probabilistically assign individual photons to sources, a key distinction when analyzing low-count X-ray data.

### 2.1.1 Statistical approach

Here we use finite mixture distributions to model several overlapping sources of photons in a high-energy image. Finite mixture distributions are a useful class of statistical models for data that are drawn from a mixture of several subpopulations; these models are finite in that the (possibly unknown) number of subpopulations is a finite positive integer. (See [McLachlan and Peel 2004](#) and [Titterton et al. 1985](#) for comprehensive discussion of finite mixture distributions, and, for example, [Mukherjee et al. 1998](#) for a previous application in astronomy.) We take a Bayesian perspective that allows joint inference for the parameters that describe the photon sources (e.g., their number, intensities and locations), the basic shape

of their spectra, and the probability that any particular photon originated from each source, given its recorded location and energy.

Performing inference jointly on the image and spectra improves the precision of the fitted parameters, and also provides more coherent measures of uncertainty than would be available if the spatial and spectral data were analyzed separately. Furthermore, unlike other methods for overlapping sources, our approach quantifies uncertainty about the number of sources. Whether we are ultimately interested in spatial or spectral aspects of sources, identifying the correct number of sources is clearly fundamental. Consequently, a coherent measure of the uncertainty associated with the fitted number of sources is critical to the appropriate interpretation of the fitted parameters of the individual sources.

In some applications inference for the number of sources may seem unnecessary because the sources are clearly identifiable. For instance, the *XMM-Newton* observation of FK Aqr and FL Aqr analyzed in Section 2.6 has relatively weak background noise, and the sources overlap only moderately. In such cases, the main advantage of the proposed method is that it precisely quantifies the uncertainties associated with the positions, intensities and spectral shapes of the sources. As already mentioned, finite mixture analysis also yields, for each observed photon, the probability that it originated from each inferred source (or the background). In this way we do not deterministically assign photons to sources, but rather properly assess the uncertainty of their origins. This is in contrast to other methods, such as those based on source regions, which deterministically assign photons to nearby sources, and therefore do not properly quantify uncer-

tainties in fitted source parameters.

There is a potential for overfitting in finite mixture models if the number of sources is unknown. This is mitigated when substantial prior information regarding the shape of the PSF or the number of sources is available, or both. In practice, we have detailed information about the PSF, and hence know exactly what the distribution of the recorded photon locations should be for each source. (For point sources this is trivial, but even for extended sources one can easily convolve the source model with the PSF.) Even if the PSF varies across the field, the shape of the photon scatter is completely determined by the location of the source. With this complete knowledge of the PSF, there is only a small risk of overfitting, even with limited prior information regarding the number of sources and their spectral shapes. Indeed, our results do not strongly depend on the choice of prior distribution for the number of sources (see Section 2.5.1).

Our method is designed for analyzing images composed of an unknown number of point sources that are contaminated with background. However, it can be applied to extended sources, with some modifications to account for spatial variations in intensity and spectra. We also mention that the success of our method depends partly on our ability to use spectra to distinguish point sources from the background, which is possible because a typical X-ray point source spectrum is more peaked than the background. Because of this, we are able to use basic models that capture the rough spectral shape in order to exploit spectral information whilst conserving computational resources. In the X-ray band, this approach offers substantial improvements over analyses using only spatial data without the

cost of precisely modeling the spectra. However, the utility of the method in other wavelength bands will depend somewhat on the nature of the spectra typical of those bands.

The remainder of the paper is organized into seven sections. Section 2.2 develops the statistical model for isolated sources in the context of high-energy datasets, and describes how these models are combined in the case of multiple sources. Section 2.3 uses a motivating example to illustrate the method and the benefits of incorporating spectral models for the sources. The beginning of Section 2.4 gives a brief review of Bayesian inference. The remainder of Section 2.4 describes the details of the proposed Bayesian analysis and computational approach. Section 2.5 presents two simulation studies. The first illustrates that inference for the number of sources is insensitive to the choice of prior distribution, and the second more thoroughly studies the advantages of using the spectral data. Sections 2.6 and 2.7 present the results of our analysis of observations from the *XMM-Newton* and *Chandra* X-ray observatories. The *XMM* observation is of the apparent visual binary FK Aqr and FL Aqr, and the *Chandra* observation is of approximately 14 sources from near the center of the Orion nebula. We summarize in Section 2.8 and computational details are in the appendices. Our *Bayesian Separation of Close Sources (BASCS)* software is available on GitHub at <https://github.com/astrostat/BASCS>.

## 2.2 Data and statistical models

### 2.2.1 Structure of the data

High-energy detectors record directional coordinates  $(x_i, y_i)$  and energy  $E_i$  for each detected photon, where  $i = 1, \dots, n$  indexes the photons. As mentioned, in practice, the PI channel is used to quantify energy. We denote the full set of spatial and spectral information for  $n$  detected photons by  $(x, y, E)$ . These observed quantities are subject to the effects of the PSF and the spectral Redistribution Matrix Function (RMF). We explicitly account for PSF effects in our model, but model the *observed* spectra, the convolution of the source spectra and the RMF. This strategy does not allow us to fit source spectral models, but does allow us to leverage spectral data to separate the sources. Even though all the attributes are recorded digitally and are binned quantities, we treat them as continuous variables for simplicity, since this binning is at scales that heavily over-sample the PSF.

Each photon is assumed to originate from one of several point sources or the background, but its exact origin is unknown. Furthermore, the number of point sources contributing photons to the data, their locations, intensities, and spectral distributions are all unknown. We assume background is distributed uniformly across the image, its strength and spectral distribution are often not known.

### 2.2.2 Prototype model for a single source

To introduce notation and our model in the simplest case, we first suppose that the data consist of photons from a single source, with no background contamination.



Statistical models specify a distribution for the observed data conditional on a number of typically unknown parameters; we discuss parameter fitting Section 2.4.3. In the current case, given the unknown position of the source, the detected photons are assumed to be dispersed according to a PSF. That is,

$$(x_i, y_i) | \mu \sim \text{PSF centered at } \mu \quad (2.1)$$

for  $i = 1, \dots, n$ , where  $\mu = (\mu_x, \mu_y)$  is the unknown position of the source<sup>2</sup>. We use the same 2-D King profile<sup>3</sup> in all the simulations and data analyses presented, see [Read et al. \(2010\)](#) and [King \(1962\)](#). The King profile density, shown in Figure A.2 in Appendix A.4, has heavy tails and is essentially a bivariate Cauchy distribution. Specific parameter values are detailed in Appendix A.4. More generally, although our method assumes that the PSF is known given  $\mu$ , it may vary with  $\mu$ . Furthermore, the PSF may be any function which can be quickly evaluated analytically or numerically. Even in cases where computationally expensive evaluations are required our method is feasible if the PSF is first tabulated.

An important feature of our overall approach is that it also utilizes the spectral data to better assess the likely origin of each photon (when background or more than one source is present). With this end in mind, we propose a simple and computationally practical model for the basic shape of the source spectrum. In

---

<sup>2</sup>The notation  $x|z \sim F$  means that, the variable  $x$  has the distribution denoted by  $F$  if  $z$  is fixed and known, and we say that  $x$ , given  $z$ , follows the distribution  $F$ . Throughout, when we use this notation we mean that repeated realizations of  $x$  are independent given  $z$ .

<sup>3</sup>The `beta2d` model in `CIAO/Sherpa`.

particular, we model photon energies using a *gamma* distribution,<sup>4</sup>

$$E_i|\alpha, \gamma \sim \text{gamma}(\alpha, \alpha/\gamma) \quad (2.2)$$

for  $i = 1, \dots, n$ . Here,  $\alpha$  and  $\gamma$  are the unknown shape and mean parameters used to describe the basic spectral distribution.<sup>5</sup> The *gamma* distribution allows flexible modeling of positive quantities with right skewed distributions.<sup>6</sup> We emphasize that we aim to summarize the essential shape of the spectral distribution, rather than to model the details of emission lines and other spectral features. This is practical because for high-energy missions, the effective areas are typically small at low and high energies, with a broad peak in the middle; the resulting counts spectrum is reasonably modeled by a single- or double-component *gamma* distribution (particularly since we ignore the RMF). Our goal is to identify sources and divide photons among them, not to carry out detailed spectral analysis. However, our algorithm allows for complex spectral models to be built in if necessary. In addition, and computationally more feasible, once the *gamma* model has fulfilled its role in separating sources, a more sophisticated spectral model may then be used to draw scientific conclusions about the spectral distributions of the disentangled sources. This final stage will be discussed in Section 2.7.2.

---

<sup>4</sup>A standard parameterization of the  $\text{gamma}(\alpha, \beta)$  distribution yields the density  $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \frac{\alpha^\alpha}{\gamma^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{\alpha}{\gamma} x}$ ,  $x > 0$ . Here  $\alpha$  and  $\beta = \alpha/\gamma$  are the shape and rate parameters, respectively.

<sup>5</sup>We parameterize the *gamma* distribution using the shape and mean, instead of the shape and rate, for interpretability and because computationally it is best to avoid rates, which in our applications tend to be close to the parameter space boundary at zero.

<sup>6</sup>Indeed, the Exponential and Chi-squared distributions are special cases, and a *gamma* can also closely resemble a (truncated) Gaussian distribution.

### 2.2.3 Prototype model for multiple sources

In practice there are multiple sources and background contamination, hence we introduce a *finite mixture model*. Let  $K$  be a parameter denoting the number of sources and  $\mu = (\mu_1, \dots, \mu_K)$  be a  $2 \times K$  matrix giving the source positions i.e.  $\mu_j = (\mu_{jx}, \mu_{jy})$ , for  $j = 1, \dots, K$ . If we knew the origin of every photon then, we could model the spatial and spectral data associated with each point source as we did in Section 2.2.2. We thus introduce a new variable  $s_i$  which indicates the source number associated with photon  $i$ . Each  $s_i$  takes on a value between 1 and  $K$ , and we let  $s$  denote the vector  $(s_1, \dots, s_n)$ . Note that  $s_i$  is never actually observed and thus is a *latent variable*. A latent variable is essentially an unknown parameter which is useful for modeling, but may not be of direct interest in itself. Here, we have introduced  $s_i$  to simplify the model and to facilitate the algorithms used for inference, which are described in Section 2.4.3.

As a parameter,  $s_i$  is also conditioned on in our spatial model, which now becomes

$$(x_i, y_i) | (\mu, s_i = j) \sim \text{PSF centered at } \mu_j \quad (2.3)$$

for  $i = 1, \dots, n$ . As an unknown parameter,  $s_i$ , plays a role similar to  $\mu$ ; it is “given” in (2.3). The spectral model can also be straightforwardly generalized to the multiple source case. We have

$$E_i | (\alpha_j, \gamma_j, s_i = j) \sim \text{gamma}(\alpha_j, \alpha_j / \gamma_j) \quad (2.4)$$

for  $i = 1, \dots, n$ , where the parameters  $\alpha_j$  and  $\gamma_j$  usually differ among the sources.

In addition to point sources, we must model the background. To this end we extend the set of possible values of  $s_i$  to include 0. Throughout, symbols indexed by 0 refer to the background. We assume that photons originating from the background are uniformly distributed across the image,

$$(x_i, y_i) | (\mu, s_i = 0) \sim \text{Uniform} \quad (2.5)$$

for  $i$  such that  $s_i = 0$ . Instrument effects may cause the background to be non-uniform, and a refinement would be to model such effects.

The background spectrum is also assumed to be flat over the energy range of the source spectra. That is, it is assumed to have a uniform distribution on  $(E_{\min}, E_{\max})$ , where  $E_{\min}$  and  $E_{\max}$  are the minimum and maximum photon energy observed. This is a good approximation because the background spectrum is expected to be less peaked than that of a point source.

So far we have not considered the intensities of the different sources and the background. Naturally there should be a parameter for each source, and one for the background, to specify the intensities. Let  $n_j$  denote the number of photons originating from source  $j$ , for  $j = 0, \dots, K$  (with zero denoting the background), mathematically,<sup>7</sup>  $n_j = \sum_{i=1}^n I_{\{s_i=j\}}$ . We can realistically model  $n_j$  as a Poisson variable with some mean  $m_j$ , for  $j = 0, \dots, K$ . Because these Poisson means vary with exposure time, however, the relative intensities,  $w_j = m_j / \sum m_j$ , are of more direct interest. Writing  $w = (w_0, \dots, w_K)$ , and given  $n$ , the Poisson model for

---

<sup>7</sup> $I$  is an indicator function that is zero if its argument is false and one otherwise.

$(n_0, \dots, n_K)$  yields a Multinomial model,<sup>8</sup>

$$(n_0, \dots, n_K) | w, n, K \sim \text{Multinomial}(n; w), \quad (2.6)$$

where  $\sum_{j=0}^K w_j = 1$ . Under this parameterization, the relative strengths of the sources and background can be succinctly expressed by the vector  $w = (w_0, \dots, w_K)$  without further reference to  $n$ . Accordingly, all inference is performed given  $n$ , because its value tells us nothing about the number of sources or their parameters.

To complete our introduction of the model we derive the likelihood function, which is the probability of the data expressed as a function of the parameters. The likelihood tells us what values of the parameters are supported by the data and is a key component for principled statistical inference. Let  $\mathcal{I}_j$  be the set of photons originating from source  $j$  (including  $j = 0$ ) and let  $\mathcal{I}$  be the entire collection of observed photons.<sup>9</sup> Also, denote the value of the PSF centered at  $\mu$  and evaluated at  $(x, y)$  by  $f_\mu(x, y)$ . Lastly, here and throughout, we let  $\theta_j = \{w_j, \mu_j, \alpha_j, \gamma_j\}$  denote the parameters associated with source  $j$ , for  $j = 1, \dots, K$ . Similarly, for the background, we let  $\theta_0 = w_0$ . We let  $\Theta_K$  denote all the source (and background) specific parameters i.e.  $\Theta_K = \{\theta_0, \dots, \theta_K\}$ . The remaining parameters are  $K$  and  $s$ . As already discussed, we treat  $n$  as fixed, and impose the constraint that the likelihood is zero unless  $\sum_{j=0}^K n_j = n$ . Combining the different parts of the model

---

<sup>8</sup>The Multinomial distribution assigns the probability  $(n!/(n_0! \dots n_K!)) \prod w_0^{n_0} \dots w_K^{n_K}$  to the allocation given by  $(n_0, \dots, n_K)$  of  $n = \sum_{i=0}^K n_i$  objects into  $K + 1$  categories.

<sup>9</sup>Mathematically,  $\mathcal{I}_j$  is the set of photon indices associated with source  $j$ , that is,  $\mathcal{I}_j = \{i : s_i = j\}$ , for  $j = 0, \dots, K$ , and  $\mathcal{I} = \bigcup_{j=0}^K \mathcal{I}_j = \{1, \dots, n\}$ .

yields the *full model* likelihood

$$L_n^{\text{full}}(\Theta_K, K) \equiv p(x, y, E | \Theta_K, K, s, n) \propto \prod_{i \in \mathcal{I}/\mathcal{I}_0} f_{\mu_{s_i}}(x_i, y_i) g_{\alpha_{s_i}, \gamma_{s_i}}(E_i), \quad (2.7)$$

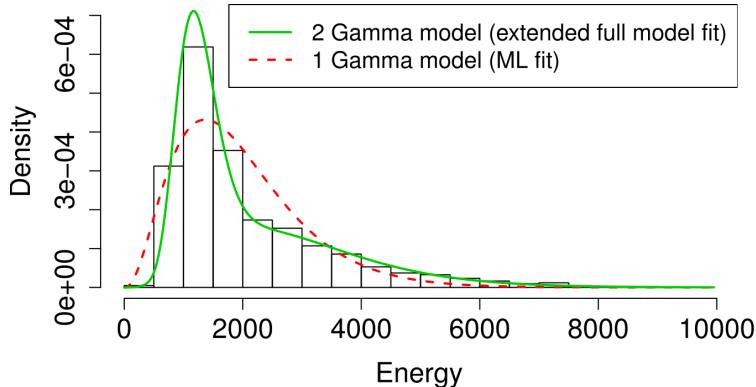
where

$$g_{\alpha_{s_i}, \gamma_{s_i}}(E_i) = \frac{\alpha_{s_i}^{\alpha_{s_i}}}{\gamma_{s_i}^{\alpha_{s_i}} \Gamma(\alpha_{s_i})} E_i^{\alpha_{s_i} - 1} e^{-\alpha_{s_i} E_i / \gamma_{s_i}}. \quad (2.8)$$

The maximum energy  $E_{\text{max}}$  and the image area are assumed to be known quantities, rather than parameters to be inferred. They are therefore omitted from the likelihood, as are all terms not involving the parameters. In later sections, we compare analyses under the full model to analyses under the *spatial-only model* that does not use the spectral information. The likelihood of the spatial-only model is

$$L_n^{\text{sp}}(\Theta_K^{\text{sp}}, K) \equiv p(x, y | \Theta_K^{\text{sp}}, K, s, n) \propto \prod_{i \in \mathcal{I}/\mathcal{I}_0} f_{\mu_{s_i}}(x_i, y_i). \quad (2.9)$$

The notation  $\Theta_K^{\text{sp}} = \{w_0, \dots, w_K; \mu_1, \dots, \mu_K\}$  represents the set of spatial parameters. Note that, although  $w$  does not explicitly appear in either likelihood, the data does nevertheless constrain  $w$  in both cases. In particular, the likelihoods indicate probable values of  $s$  which in turn indicate probable values of  $w$ . Conceptually, our method is to apply Bayes rule, briefly reviewed in Section 2.4.1, to the likelihoods displayed in (2.7) and (2.9) to yield a distribution summarizing our knowledge of the parameters given the data, i.e. the joint posterior distribution.



**Figure 2.1:** Fitting *gamma* distributions to a counts spectrum. The histogram shows the observed spectrum of the brightest of the *Chandra* sources in the Orion field in Section 2.7.2 (from one iteration of our algorithm; see Section 2.4.3), and the curves show *gamma* model fits. The solid line (green) is the extended full model fit of the two-*gamma* spectral model and the dashed line (red) is the maximum likelihood fit of the one-*gamma* model.

## 2.2.4 Extensions of the spectral model

In some situations the *gamma* spectral model given by (2.4) is not sufficiently flexible to capture the spectral shape of the observed sources. For example, Figure 2.1 shows the observed spectrum of the brightest source in the *Chandra* observation analysed in Section 2.7. In particular, the histogram shows the spectrum using one likely assignment of photons produced during the iterations of our algorithm (see Section 2.4.3). The dashed red curve shows the maximum likelihood fit of the *gamma* distribution to the observed spectrum. The *gamma* does not fit the distribution closely. This causes a problem because inference based on the (mis-specified) *gamma* spectral model will suggest there are two sources instead of one in order to better capture the spectral distribution of the source.

To solve this problem, we use a mixture of two *gamma* distributions for a more

general spectral model. That is,

$$E_i | (\alpha_{j1}, \alpha_{j2}, \gamma_{j1}, \gamma_{j2}, \pi_{j1}, \pi_{j2}, s_i = j) \sim \sum_{l=1}^2 \pi_{jl} \text{gamma} \left( \alpha_{jl}, \frac{\alpha_{jl}}{\gamma_{jl}} \right), \quad (2.10)$$

for  $i = 1, \dots, n$ , where the parameters  $\pi_j$  and  $\pi_{j2} = 1 - \pi_{j1}$  are the weights of the two *gamma* components. When this two *gamma* mixture spectral model is substituted for the one *gamma* spectral model in (2.7) we obtain the following *extended full model* likelihood

$$\begin{aligned} L_n^{\text{ext}}(\Theta_K^{\text{ext}}, K) &\equiv p(x, y, E | \Theta_K^{\text{ext}}, K, s, n) \\ &\propto \prod_{i \in \mathcal{I}/\mathcal{I}_0} \left( f_{\mu_{s_i}}(x_i, y_i) \sum_{l=1}^2 \pi_{s_i l} g_{\alpha_{s_i l}, \gamma_{s_i l}}(E_i) \right). \end{aligned} \quad (2.11)$$

The notation  $\Theta_K^{\text{ext}}$  denotes  $\{\theta_0^{\text{ext}}, \dots, \theta_K^{\text{ext}}\}$ , where  $\theta_j^{\text{ext}} = \{w_j, \mu_j, \alpha_{j1}, \alpha_{j2}, \gamma_{j1}, \gamma_{j2}, \pi_{j1}, \pi_{j2}\}$  gives the parameters associated with source  $j$ , for  $j = 1, \dots, K$ , and  $\theta_0^{\text{ext}} = \theta_0$ . The solid green curve in Figure 2.1 shows the extended full model fit of the *gamma* mixture spectral model. In this example, the mixture of *gammas* quite closely fits the observed spectrum and generally there did not appear to be unwarranted splitting of sources into two in our numerical studies using this model.

Even greater flexibility of the spectral model could be gained by considering a mixture of more than two *gammas*, but this was not necessary in our numerical studies. For the *XMM* data of Section 2.6, the one-*gamma* spectral model is sufficient in that, for both of the sources, the maximum likelihood fit of the one-



*gamma* and the two-*gamma* models resulted in essentially identical fits when using a feasible allocation of photons. In the interest of simplicity, we only use the extended full model when necessary (i.e., in Section 2.7), and elsewhere use the full model given in (2.7).

### **Detecting spectral model inadequacy**

A natural question is how one should decide if the source spectral model is inadequate for our purpose of allocating photons among the different sources (and background). There are two potential indications of spectral model misspecification. Firstly, analysis may tend to divide bright sources into two. In particular, when the algorithm (see Section 2.4.3) finds many instances of sources very close together this indicates that the spectral model is probably not adequate.<sup>10</sup> A second indication of inadequacy of the spectral model comes from considering inference under the spatial-only model. We can inspect the empirical distribution of the photons assigned to a source in iterations of the spatial-only algorithm. If this empirical distribution differs substantially from a *gamma* distribution then it is unlikely that the one-*gamma* spectral model is sufficiently flexible. Clearly, looking at the empirical spectral distribution of a source under the spatial-only model is only reliable if we can accurately assign photons based on spatial data alone. Thus, when possible it is best to select a bright source which is relatively isolated. In the presence of uncertainty about the shape of the spectral distributions to expect then it is usually sensible to use a mixture of at least two *gammas* (or

---

<sup>10</sup>Misspecification of the PSF, and specifically under-estimation of its width, could have a similar effect.

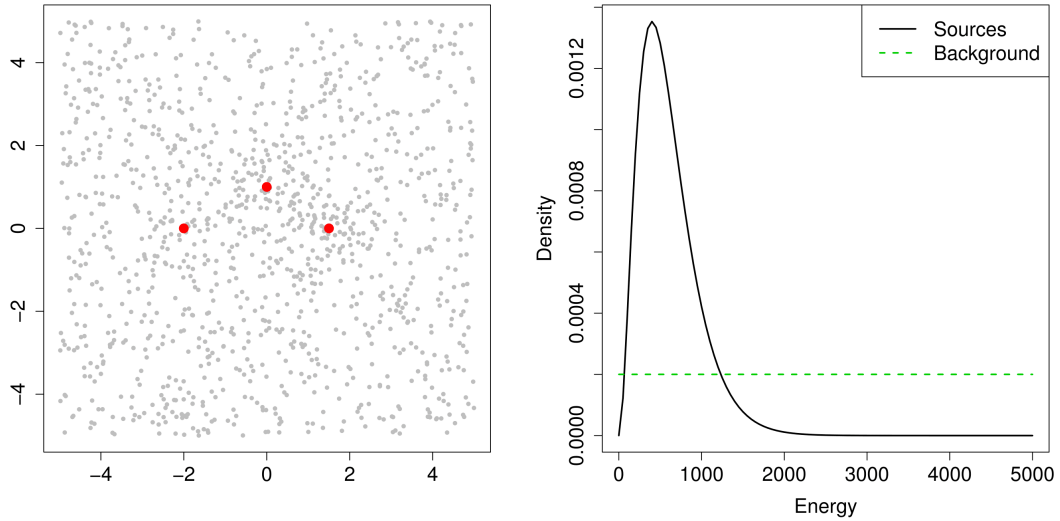
perform analysis several times using mixtures of different numbers of *gammas*). In the presence of uncertainty about the shape of the spectral distributions to expect then it is usually sensible to use a mixture of at least two *gammas* (or perform analysis several times using mixtures of different numbers of *gammas*). One should be cautious of using a spectral model that is too complicated<sup>11</sup> because overfitting may decrease the benefits of modeling the spectral data.

## 2.3 Illustrative example

To motivate our method we present a simple simulated data example that illustrates the potential gains made possible by using the full model instead of the spatial-only model. We emphasize that this is a walk-through, designed to clarify the conceptual foundations of the method. A detailed description of our method is in Section 2.4. The simulated data consist of the spatial and spectral details of photons detected from three weak sources contaminated with background. The spatial data and the spectral distributions used for simulation are shown in Figure 2.2. The background average is 10 photons per unit square, and the numbers of photons from each source are drawn from Poisson distributions with means 100, 50 and 25, respectively. Thus, the background is very strong and contributes about 85% of the photons over the entire image, and about 40%, 53%, and 66% respectively in the three source regions. The true source positions are  $(1.5, 0)$ ,  $(0, 1)$ , and  $(-2, 0)$ , and their source regions are approximately circles of radius

---

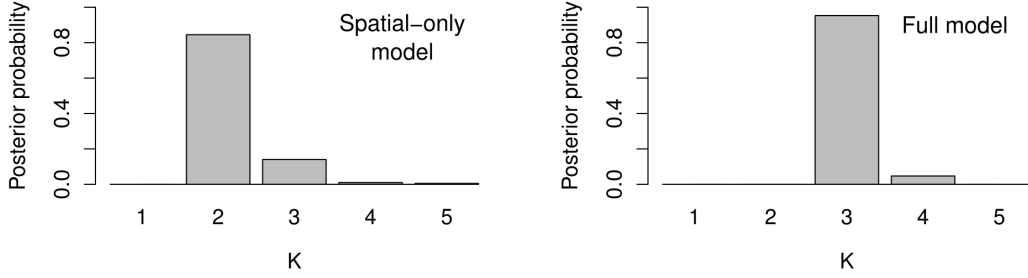
<sup>11</sup>We can avoid an overly complicated model by imposing parametric constraints or utilizing substantial prior information to be sure only scientifically plausible spectral shapes are allowed.



**Figure 2.2:** Illustrative simulation setup. Locations of three weak sources are shown as red dots over a scatter plot (left), as also are the adopted counts spectra of the sources and the background (right).

1. All three sources have the same PSF, the 2-D profile density shown in Figure A.2 in Appendix A.4. The source spectral data is drawn from a *gamma* distribution plotted in Figure 2.2 (mean parameter 600 and shape parameter 3). In this simple illustration, all the sources have the same theoretical spectral distribution; however, this is not assumed in the fitted model, which is based on the likelihood in (2.7). The theoretical background spectrum was uniform on  $(0, 5000)$ .

We fit both the spatial-only and the full model to the simulated data. The resulting posterior probability distributions for  $K$  are shown in Figure 2.3. With the spatial data alone it is difficult to detect the faintest source, and consequently the most likely value of  $K$  is 2 rather than 3. The situation is much improved when we include the spectral data. The advantage of using the spectral information is due to a greater ability to distinguish the sources from the background, owing to



**Figure 2.3:** Probability distribution of the number of sources based on the spatial-only model (left) and the full model (right). In this simulation, the true value is  $K = 3$ .

**Table 2.2:** Fitted parameters under the full and spatial-only models. The columns in bold give the fits that would likely be relied upon in practice for the two models. The intervals in parentheses indicate the 16% and 84% posterior quantiles, i.e., Bayesian  $1\sigma$  equivalent intervals.

	Truth	Full model		Spatial-only model	
$k$	3	<b>3</b>		<b>2</b>	3
$P(K = k \text{data})$	–	<b>0.95</b>		<b>0.85</b>	0.14
$\mu_{1x}$	1.5	<b>1.51</b> (1.41,1.61)		<b>1.43</b> (1.27,1.58)	1.44 (1.29,1.59)
$\mu_{1y}$	0	<b>-0.01</b> (-0.10,0.09)		<b>0.04</b> (-0.08,0.17)	0.02 (-0.10,0.14)
$\mu_{2x}$	0	<b>-0.08</b> (-0.20,0.04)		<b>-0.09</b> (-0.28,0.12)	-0.03 (-0.22,0.15)
$\mu_{2y}$	1	<b>1.11</b> (1.00,1.23)		<b>0.96</b> (0.80,1.13)	0.99 (0.84,1.15)
$\mu_{3x}$	-2	<b>-1.96</b> (-2.17,-1.76)		–	-1.37 (-2.40,0.35)
$\mu_{3y}$	0	<b>0.06</b> (-0.15,0.27)		–	-0.24 (-1.44,0.75)
$w_1$	0.083	<b>0.068</b> (0.057,0.078)		<b>0.063</b> (0.049,0.076)	0.062 (0.049,0.076)
$w_2$	0.058	<b>0.064</b> (0.053,0.076)		<b>0.055</b> (0.041,0.068)	0.052 (0.039,0.066)
$w_3$	0.033	<b>0.028</b> (0.019,0.036)		–	0.017 (0.003,0.030)
$w_0$	0.826	<b>0.841</b> (0.826,0.855)		<b>0.883</b> (0.866,0.900)	0.868 (0.848,0.887)
$\gamma_1$	600	<b>536</b> (478,592)		–	–
$\gamma_2$	600	<b>735</b> (646,820)		–	–
$\gamma_3$	600	<b>634</b> (397,826)		–	–
$\alpha_1$	3	<b>3.92</b> (2.89,4.97)		–	–
$\alpha_2$	3	<b>2.94</b> (2.18,3.69)		–	–
$\alpha_3$	3	<b>2.76</b> (1.62,3.82)		–	–

**Table 2.3:** Photon allocation proportions for the spatial-only and full models.

Source (true intensity)	No. Photons in simulation	Average allocation probabilities							
		Spatial-only model				Full model			
		Background	Right	Middle	Left	Background	Right	Middle	Left
Background (10/sq)	1001	0.917	0.037	0.033	0.013	0.940	0.022	0.026	0.012
Right (100)	84	0.566	0.354	0.068	0.012	0.318	0.557	0.113	0.012
Middle (70)	67	0.593	0.073	0.303	0.031	0.313	0.122	0.505	0.060
Left (40)	42	0.800	0.034	0.071	0.095	0.431	0.066	0.145	0.358

the difference between the source spectra and the background spectrum.

Modeling the spectral data also improves estimation of the other parameters, even if we consider the fits based on  $K = 3$ . (This is the correct value of  $K$  and is identified by the full model but not the spatial-only model.) In Table 2.2, the first bold column and the last column (not bold) show a summary of the fitted parameters for  $K = 3$  under the full model and spatial-only model, respectively. When we consider  $K = 3$ , the greatest gains of using the full model are in estimating the parameters of the faintest source because this source is the hardest to distinguish from the background when using only spatial data.

In practice, the advantage of using the spectral data for estimating the source parameters is greater than is apparent when we only consider  $K = 3$ . When confronted with the summary of the fit of  $K$  under the spatial-only model (given in the left panel of Figure 2.3), a researcher is likely to rely on the parameter fits assuming  $K = 2$ . Thus, it is fair to compare the  $K = 3$  fit under the full model with the  $K = 2$  fit under the spatial-only model (i.e., the bold columns in Table 2.2). The latter is clearly substantially worse than the former, because the faint source goes undetected and has no fitted parameters.

The improvement in separation of the sources (and background) can be further understood from Table 2.3, which summarizes the probability that each photon originated from each source or the background, again under the optimistic assumption that  $K = 3$  (see Section 2.4.3 for additional details). The rows of the table indicate the true photon origin, and the columns indicate the fitted origins. The table entries are the average probabilities, across photons, of the different

fitted origins. Ideally the matrices would be identity matrices with ‘1’s along the diagonal and ‘0’s elsewhere, but because of the strength of the background many source photons are mixed up in the background. For example, for a photon originating from the leftmost source, the spatial-only model on average assigns probabilities of 0.095 and 0.800 that it originated from the correct source and the background respectively, reflecting the difficulty in detecting the location of this faint source. Under the full model, the average probability of correct assignment is increased to 0.358, a substantial improvement. Indeed, for each of the three sources, *nearly half as many photons are mixed up with the background under the full model*. Our improved ability to correctly assign photons under the full model (relative to the spatial-only model) naturally leads to improved estimation of the parameters of the faint source, as illustrated in Table 2.2. There is a similar effect for the other sources though it is less pronounced because, being brighter, they are easier to detect from the spatial data alone.

## 2.4 Bayesian model fitting

### 2.4.1 Bayesian inference

The Bayesian perspective provides a coherent approach for combining all available information to infer the unknown model parameters  $\Theta_K$ ,  $K$ , and  $s$ . Firstly, our knowledge (or lack of knowledge) as to the likely values of the parameters before seeing the current data is quantified using a *prior distribution*. Once the data are observed, Bayes’ Theorem allows us to combine the likelihood and the prior

distribution to yield the *posterior distribution* of the parameters. Recall, the likelihood is the probability of the data given the parameters. The posterior distribution expresses our updated knowledge of the parameters after seeing the data. Bayes' Theorem states that, for generic data and parameter vector  $\psi$ , the posterior distribution is

$$p(\psi|\text{data}) = \frac{p(\text{data}|\psi)p(\psi)}{p(\text{data})}, \quad (2.12)$$

where  $p(\text{data}|\psi)$  is the likelihood function and  $p(\psi)$  is the prior distribution. The denominator  $p(\text{data})$  is simply a normalizing constant which ensures the posterior integrates to one. In our case, the data is  $(x, y, E)$  and under the full model  $\psi = \{\Theta_K, K, s\}$  so

$$p(\Theta_K, K, s|x, y, E) = \frac{p(x, y, E|\Theta_K, K, s)p(\Theta_K, K, s)}{p(x, y, E)}. \quad (2.13)$$

Here, all probabilities are conditional on  $n$  but this is suppressed. The likelihood  $p(x, y, E|\Theta_K, K, s)$  is given in (2.7), and the prior distribution  $p(\Theta_K, K, s)$  is described in Section 2.4.2. Referring back to the illustrative example in Section 2.3, the marginal posterior distribution of  $K$ ,

$$p(K|x, y, E) = \sum_s \int p(K, s, \Theta_K|x, y, E)d\Theta_K, \quad (2.14)$$

is displayed in Figure 2.3. Given the number of unknown parameters, it is not possible to plot their joint posterior distribution, but we can derive and plot the

marginal posterior distribution of any one parameter, as in (2.14) and Figure 2.3.

## 2.4.2 Completing the model formulation: prior distributions

Following the Bayesian approach, we specify prior distributions for each of the unknown parameters. Firstly, the positions of the point sources are *a priori* assumed to be independently and uniformly distributed across the image. That is,

$$\mu_j \sim \text{Uniform} \tag{2.15}$$

for  $j = 1, \dots, K$ . In principle, informative priors can be used if prior information on source locations is available. For example, we might set  $\mu_j \sim N(\mu_{j0}, \sigma_{j0}^2)$ , where  $(\mu_{j0}, \sigma_{j0})$ , for  $j = 1, \dots, K$ , specifies knowledge of the source locations.<sup>12</sup>

Next, the vector  $w$ , that specifies relative intensities, is given a Dirichlet<sup>13</sup> prior distribution with parameter  $(\lambda, \dots, \lambda)$ . A Dirichlet random variable is a probability vector, i.e., it is a vector with non-negative entries that sum to one. We set  $\lambda = 1$  throughout. This choice is uniform on the probability vector, but very slightly favors sources of equal size. Indeed, setting  $\lambda = 1$  means the Dirichlet prior has as much information as a single photon count added to each source

---

<sup>12</sup>The notation  $N(\mu_{j0}, \sigma_{j0}^2)$  denotes a Gaussian distribution with mean  $\mu_{j0}$  and variance  $\sigma_{j0}^2$ .

<sup>13</sup>The Dirichlet density is  $f(p_0, \dots, p_K) = \left( \Gamma(\sum_{i=0}^K \lambda_i) / \prod_{i=0}^K \Gamma(\lambda_i) \right) \prod_{i=0}^K p_i^{\lambda_i - 1}$ , for all  $p_i$  such that  $\sum_{i=0}^K p_i = 1$  and  $p_i \geq 0$  for  $i = 0, \dots, K$ , and is zero otherwise. Here,  $(\lambda_0, \dots, \lambda_K)$  is a parameter, and  $\Gamma$  is the gamma function.



(including a single count added to the background).<sup>14</sup> Regarding the realized vector of source and background counts  $(n_0, \dots, n_K)$ , recall that (2.6) specifies a Multinomial distribution for  $(n_0, \dots, n_K)$ , given  $w$ ,  $n$ , and  $K$ . Since  $(n_0, \dots, n_K)$  is a function of the parameter (or latent variable)  $s$ , (2.6) is effectively a prior distribution for  $s$ .<sup>15</sup>

External information about the number of sources is amalgamated into a prior for  $K$ , which we assume to be Poisson with mean parameter  $\kappa$ .<sup>16</sup> Under the Poisson prior, the fitted value of  $K$  is relatively robust to the choice of  $\kappa$  because the PSF is completely specified.<sup>17</sup> Indeed, we show in Section 2.5.1 that the posterior mode for the number of sources may correctly identify the true value of  $K$ , even when  $\kappa$  is quite different from  $K$ . Therefore, in practice it is adequate to use the Poisson prior for  $K$  with  $\kappa$  set to any reasonable guess of the number of sources.

---

<sup>14</sup>Suppose the source counts are observed to be  $(n_0, \dots, n_K)$  and follow a Multinomial distribution with probability vector  $w$ . Then, assuming *a priori*  $w \sim \text{Dirichlet}(\lambda_0, \dots, \lambda_K)$ , it can be shown that  $w|(n_0, \dots, n_K) \sim \text{Dirichlet}(n_0 + \lambda_0, \dots, n_K + \lambda_K)$ . Because  $\lambda_j$  is treated just like  $n_j$  in this posterior distribution,  $\lambda_j$  can be viewed as a “prior count” and we say the Dirichlet prior is as informative as  $\lambda_j$  counts added to source  $j$ , for  $j = 0, \dots, K$ .

<sup>15</sup>The parameter  $w$  is called a *hyper-parameter* because it appears in the prior distribution of  $s$  but is itself of interest and thus has its own prior distribution.

<sup>16</sup>While other priors for  $K$  are possible, the Poisson is simple and only moderately informative. Indeed, the equality of mean and variance captures the typical level of prior information we expect, e.g., if we suspect 10 sources, then an analysis yielding between 8 and 12 sources would seem quite reasonable, but we are unlikely to consider, say, 100 sources as a realistic possibility. Even less informative priors may sometimes be desirable, but it generally makes sense to use any reliable prior information that is available to guard against model misspecification. (Prior information about  $K$  also helps our algorithm to converge slightly more quickly.)

<sup>17</sup>If the PSF were not fully specified, it would be difficult to distinguish a few sources with a wide PSF from many sources with a narrow PSF. Thus, the fitted number of components of a general finite mixture model can be quite sensitive to the choice of prior on this parameter. Accounting for misspecified PSFs or uncertainties in their calibration is beyond the scope of this work (see [Lee et al. 2011](#) and [Xu et al. 2014](#) for possible strategies).

To complete the model specification, we must assign prior distributions for the source spectral distribution parameters  $\alpha_j$  and  $\gamma_j$ , for  $j = 1, \dots, K$ . Typically there is sufficient data to overwhelm these prior distributions. Thus, we are not overly concerned with the exact form of these priors. For concreteness, however, we mention that one set of priors we use is  $\alpha_j \sim \text{gamma}(2, 0.5)$  and  $\gamma_j \sim \text{Uniform}(E_{\min}, E_{\max})$ , for  $j = 1, \dots, K$ , where  $E_{\min}$  is the minimum observed energy.<sup>18</sup>

To summarize, our prior distribution for the full model parameters  $\Theta_K$ ,  $K$  and  $s$  is

$$\begin{aligned}
 p(\Theta_K, K, s) &= p(\mu, \alpha, \gamma, s | K, w) p(w | K) p(K) \\
 &\propto \left( \prod_{j=0}^K \alpha_j e^{-0.5\alpha_j} \right) \prod_{j=0}^K w_j^{n_j} \left( \prod_{j=0}^K w_j \right)^{\lambda-1} \frac{\kappa^K}{K!}, \quad (2.16)
 \end{aligned}$$

where  $\mu$ ,  $\alpha$  and  $\gamma$  denote  $(\mu_1, \dots, \mu_K)$ ,  $(\alpha_1, \dots, \alpha_K)$ , and  $(\gamma_1, \dots, \gamma_K)$ , respectively. The second term on the second line of (2.16) comes from the Multinomial prior distribution for  $s$ . In the case of the extended full model given in (2.11), the priors for  $\alpha_{jl}, \gamma_{jl}$ ,  $l = 1, 2$ , are the same as those for  $\alpha_j, \gamma_j$ , and the prior for  $\pi_{j1}$  is a Beta(2, 2) distribution,<sup>19</sup> for  $j = 1, \dots, K$ . (No prior for  $\pi_{j2}$  is needed because this parameter is determined by  $\pi_{j1}$ , for  $j = 1, \dots, K$ .) The prior for the spatial

---

<sup>18</sup>More generally, if  $K$  is large and some of the sources are faint, it may be beneficial to model the distribution of the spectral parameters across the sources. This strategy is known as hierarchical modeling and is known to have statistical advantages in terms of the estimates of the individual spectral parameters. Such hierarchical spectral structures are left as a topic for future work.

<sup>19</sup>For  $\alpha, \beta > 0$ , the Beta( $\alpha, \beta$ ) distribution density is  $f(x) = (\Gamma(\alpha + \beta)/\Gamma(\alpha)\Gamma(\beta)) x^{\alpha-1}(1-x)^{\beta-1}$  for  $x \in [0, 1]$ , and is zero otherwise. Here,  $\Gamma$  is the Gamma function.

model parameters is (2.16) without the first term.

### 2.4.3 Statistical computation and model fitting

Given the likelihood in (2.7) and the prior distribution in (2.16), we can apply Bayes' Theorem to obtain the posterior distribution of  $\Theta_K$ ,  $K$  and  $s$  (see (2.13)). The resulting posterior distribution is a complicated function, which we summarize by the low-dimensional marginal distributions as described in Section 2.4.1 and their means and standard deviations. These summaries are used to estimate the model parameters and their error bars.

We accomplish the necessary numerical integration, e.g., as in (2.14), using Monte Carlo methods, a cornerstone of statistical computing (Shao and Ibrahim 2000, Liu 2008, Brooks et al. 2011). The idea of Monte Carlo algorithms is to simulate values of the generic parameter  $\psi$  from the posterior distribution in (2.12) to obtain a *Monte Carlo sample*  $\{\psi^{(1)}, \dots, \psi^{(T)}\}$ . For example, in Figure 2.3, the height of the bin centered at  $k$  is the proportion of the Monte Carlo draws with  $K^{(t)}$  equal to  $k$ , i.e.,

$$P(K = k|x, y, E) \approx \frac{1}{T} \sum_{t=1}^T I_{\{K^{(t)}=k\}}, \quad (2.17)$$

for  $k = 1, \dots, K$ .

A somewhat unusual feature of our model is that the number of parameters is determined by the value of  $K$ , the unknown number of sources. This necessarily conditional structure means that it only makes sense to consider the posterior

distributions of the other parameters for a given inferred value of  $K$  (Park et al. 2008 discuss a somewhat similar conditional inference in the context of locating emission lines). For an illustration of why this is so, consider the intensity  $w_3$  of the ‘third’ source in an image. The parameter  $w_3$  does not have the same interpretation when there are three sources versus four, because what is the ‘third’ source in the first scenario may combine two sources from the latter scenario. In fact, for  $K = 2$  the parameter  $w_3$  does not even exist. In general, there is no clear relationship between the parameters under scenarios with different values of  $K$ . This prevents us from considering the unconditional posterior distribution of, say,  $w_3$ . Instead, we are interested in posterior summaries given a particular value of  $K$ , such as  $p(w_3|K = k, x, y, E)$ . For example, the second row of Table 2.2 provides an estimate of the posterior mean of  $w_2$  conditional on  $K = 3$ , under the full model,<sup>20</sup>

$$\hat{w}_2^F(k) = \frac{\sum_{t=1}^T w_2^{(t)} I_{\{K^{(t)}=k\}}}{\sum_{t=1}^T I_{\{K^{(t)}=k\}}} = 0.080. \quad (2.18)$$

More generally, for each one-dimensional parameter  $\tau$ , we calculate the Monte Carlo estimate

$$\hat{\tau}(k) = \frac{\sum_{t=1}^T \tau^{(t)} I_{\{K^{(t)}=k\}}}{\sum_{t=1}^T I_{\{K^{(t)}=k\}}}. \quad (2.19)$$

In practice, we choose a value of  $k$  at which  $K$  has relatively high posterior probability, such as the posterior mode, because otherwise the parameters estimated

---

<sup>20</sup>The superscript  $F$  in (2.18) indicates that the Monte Carlo samples were drawn from the posterior derived under the full model.

are unlikely to correspond to properties of real sources. (Indeed, our algorithm does not accurately estimate parameters under unlikely values of  $K$ .) We may decide to consider several different values if the posterior of  $K$  is not concentrated on one value. This can be useful despite the fact that, as we have mentioned, the number and interpretation of the parameters is not consistent across values of  $K$ .

The most popular method for obtaining the Monte Carlo samples needed for estimates such as that in (2.18) is Markov chain Monte Carlo (MCMC). This an iterative algorithm in which we generate a new value of the parameters  $\psi^{(t)}$  at each iteration by drawing from a distribution  $\mathcal{F}$  that only depends on  $\psi^{(t-1)}$  (and the data) and not earlier members of the Monte Carlo sample. Continuing for  $T$  iterations we obtain a sample  $\{\psi^{(1)}, \dots, \psi^{(T)}\}$  of correlated parameter values, which is usually called an MCMC chain. Appropriate choice of  $\mathcal{F}$  ensures that the sample mimics the posterior distribution in the sense that as  $T \rightarrow \infty$  the sample empirical distribution approaches the posterior distribution. In implementation, a draw from an appropriate  $\mathcal{F}$  is typically achieved through two steps: firstly a new value of the parameters  $\psi^*$  is proposed, and then this value is either accepted or rejected with some probability.<sup>21</sup> The Metropolis-Hastings algorithm ([Metropolis et al. 1953](#) and [Hastings 1970](#)) is an example of such an algorithm. The reader is referred to [Gelman et al. \(2013\)](#) for details, including discussion of efficient choices of  $\mathcal{F}$  and monitoring of convergence to the posterior distribution (which is usually done by running multiple MCMC chains in parallel and checking that

---

<sup>21</sup>An appropriate choice of  $\mathcal{F}$  and the corresponding rejection probability to use, to ensure convergence of the sample empirical distribution to the posterior, can be calculated by appealing to the ‘reversibility condition’ (see texts on the theory of Markov chain convergence e.g. [Feller \(1968\)](#)).

their behaviour is sufficiently similar based on some criterion).

In standard MCMC algorithms the parameter space being explored is fixed throughout. In our context this means the number of sources would have to be known. We therefore turn to reversible jump Markov chain Monte Carlo (RJMCMC) algorithms (first introduced by [Green 1995](#)), which allow configurations with differing numbers of sources to be explored. There have been a number of uses of RJMCMC in other astronomy contexts, for example, [Umstätter et al. 2005](#), [Brewer and Stello 2009](#), [Jasche and Wandelt 2013](#), and [Walmiswell et al. 2013](#). In RJMCMC algorithms, so called ‘jump’ steps update the value of  $K$ , the name referring to a jump between parameter spaces (or ‘models’). These steps are performed by drawing  $K^{(t)}$  from a distribution only depending on  $\psi^{(t-1)} = (\Theta^{(t-1)}, K^{(t-1)}, s^{(t-1)})$ , in the same spirit as ordinary MCMC iterations. Feasible values of the parameters  $\Theta^{(t)}$  and  $s^{(t)}$  must simultaneously be drawn because their dimension and interpretation change with  $K$ . It is this high dimensional sampling that makes RJMCMC challenging. In RJMCMC algorithms,  $K^{(t)}$  is only allowed to differ from  $K^{(t-1)}$  by at most one. This constraint facilitates the proposal of appropriate parameters  $\Theta^{(t)}$  and  $s^{(t)}$ ; RJMCMC moves between configurations by splitting, combining, creating or destroying sources in the model. The standard RJMCMC algorithm for Gaussian mixtures was introduced in [Richardson and Green \(1997\)](#), and [Wiper et al. \(2001\)](#) illustrated RJMCMC for *gamma* mixtures. Our BASCS software essentially combines these two algorithms. Additional details are given in Appendices A.2 and A.3. For the analyses found in Sections 2.5 and 2.7 we specify the number of iterations for which our RJMCMC

algorithm was run (which depended on the observed convergence rate and run time). A single iteration of our RJMCMC algorithm consists of one proposal to change  $K$  and ten MCMC updates of the other parameters, i.e., the number of MCMC iterations is ten times greater than the stated number of RJMCMC iterations. In Section 2.6 we fix  $K$  and use MCMC, and thus directly specify the number of MCMC iterations. Our standard approach is to run ten RJMCMC (or MCMC) chains to allow monitoring of convergence, but for simplicity the final results are always computed using a single chain.

As discussed in Section 2.4.2, having detailed information about the PSF means our estimates are insensitive to the prior on  $K$  (see also Section 2.5.1). Knowledge of the PSF also aids computation in that it limits the number of feasible configurations, meaning the RJMCMC algorithm does not have to jump across many values of  $K$ . This keeps the number of iterations until approximate convergence comparatively low. Thus, knowledge of the PSF means that, despite the difficulties that are commonly thought to surround mixture models fit with RJMCMC algorithms, our proposed approach is relatively stable and robust. Nonetheless, when the number of sources is clear, MCMC algorithms should be used because they are computationally preferable to RJMCMC algorithms (see Section 2.6 for an analysis using an MCMC algorithm). In particular, MCMC algorithms are faster per iteration and fewer iterations are needed to obtain enough samples for a given  $K$  value of interest. One further challenge is moderate sensitivity to the spectral model, which is the reason why in some applications the *gamma* spectral model must be replaced by the *gamma* mixture spectral model introduced in

Section 2.2.4.

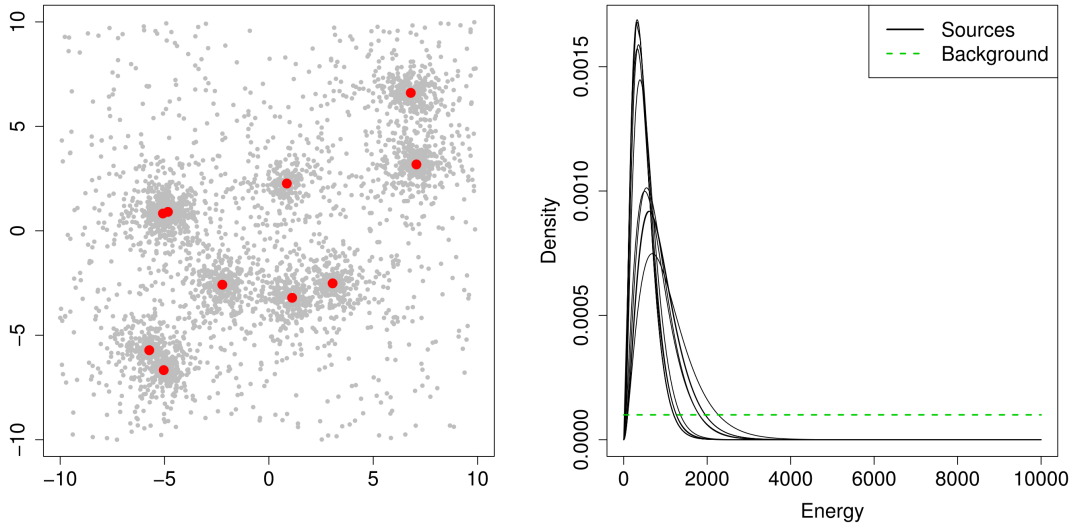
## 2.5 Simulation studies

Simulated data are used to assess two important aspects of our method: (i) the sensitivity of the fit for  $K$  on its prior distribution; and (ii) the performance of the method under a range of different source and background parameters. In the second case, of particular interest is the comparison of inference for the parameters under the spatial-only model and full models (given in (2.7) and (2.9)).

### 2.5.1 Sensitivity to prior distribution on $K$

To illustrate robustness to the prior on  $K$ , we simulated data for a one-source ( $K_{\text{true}} = 1$ ) and a ten-source ( $K_{\text{true}} = 10$ ) reality and drew inference for the number of sources under three different settings of the prior mean  $\kappa$  (1, 3, and 10). Ten datasets were simulated under each reality, each consisting of images of 20 by 20 spatial units and spectral data (simulated under the single *gamma* spectral model). We randomly placed the sources in the central 18 by 18 region of the image, avoiding the edges so that source photons are largely contained within the image. The mean number of photons  $m_j$  from source  $j$  was chosen randomly from the interval 100 to 500, for  $j = 1, \dots, K$ . The mean total number of photons from the background in each dataset,  $m_0$ , was set to 400, an average of 1 photon per unit square. The number of photons from source  $j$  (or the background) was then simulated from a Poisson distribution with mean  $m_j$ , for  $j = 0, \dots, K$ .





**Figure 2.4:** Simulated dataset for the 10 source case. The simulated spatial counts distribution (left) and the adopted spectra for each source and the background (right) are shown. The true locations of the 10 sources are marked by large (red) dots in the left plot.

Spatial coordinates for the photons were chosen by sampling from the PSF (or the Uniform distribution in the case of the background). We used the King profile density for the PSF; the same PSF is used for analysis of the datasets in Section 2.6 and Section 2.7.

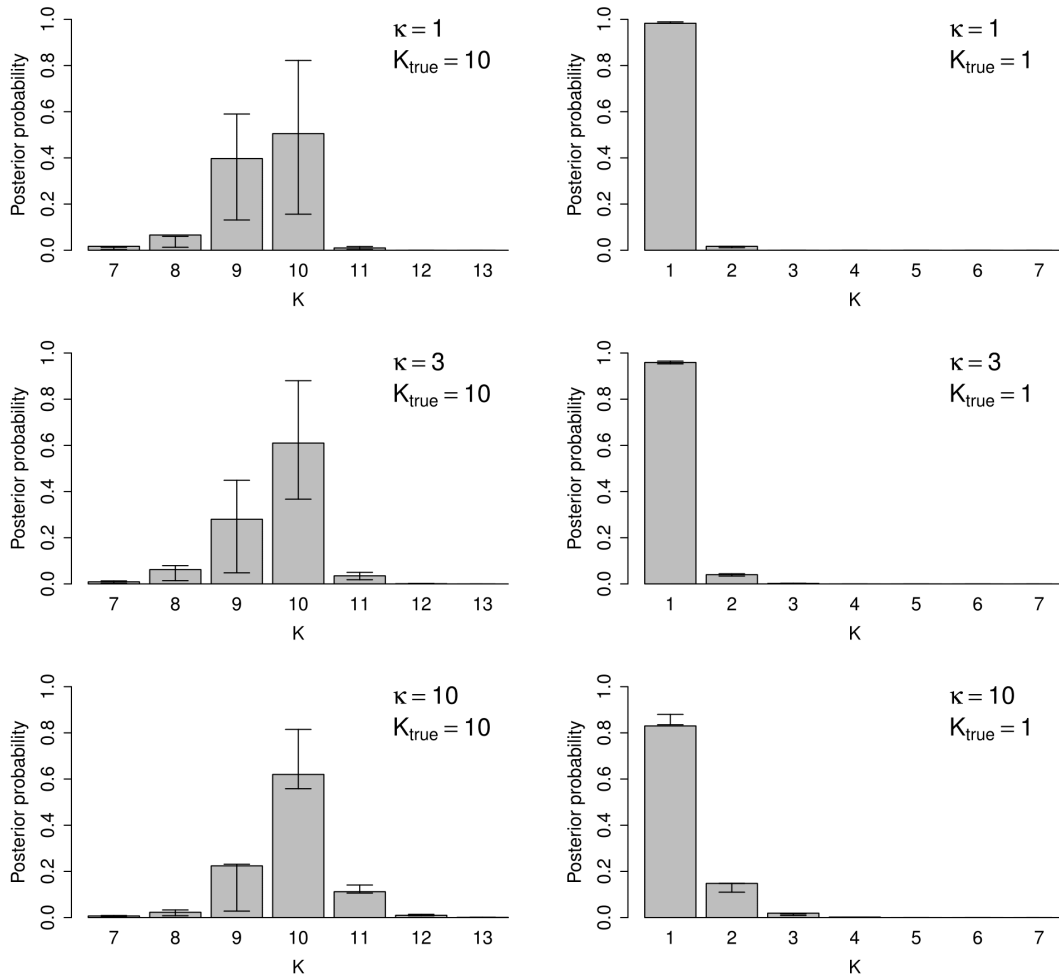
To complete the datasets, we simulated spectral data under the single *gamma* spectral model (and from a Uniform distribution in the case of the background). We drew the spectral distribution parameters  $\alpha_j$  (shape) and  $\alpha_j/\gamma_j$  (rate),  $j = 1, \dots, K$ , from the Gaussian distributions  $N(3, 0.2^2)$  and  $N(0.005, 0.001^2)$  respectively, truncating both distributions to be strictly positive. The resulting spectral parameters are similar to those fitted for the *XMM* dataset in Section 2.6. An example simulated dataset is shown in the left panel of Figure 2.4. The right panel shows the true spectral distributions for the same dataset.

For each of the 20 simulated datasets, ten RJMCMC chains were run to assess convergence, but for simplicity only one chain per dataset was used in the final analysis.<sup>22</sup> The chains were run for 200,000 RJMCMC iterations, the first 100,000 of which formed the convergence period (or burnin) and were discarded. For each dataset, the posterior probability of being in state  $K = k$  was calculated, using (2.17), for all feasible values of  $k$ . Figure 2.5 summarizes the inference for  $K$  under the ten-source ( $K_{\text{true}} = 10$ , left panels) and one-source ( $K_{\text{true}} = 1$ , right panels) realities, for  $\kappa = 1, 3$  and 10 (top, middle and bottom panels respectively). Recall that  $\kappa$  is the prior mean number of sources. The 25% and 75% quantiles of the posterior probabilities across the ten datasets are indicated for each value of  $K$ .

Figure 2.5 shows that, for the ten-source reality, the posterior probability is concentrated around  $K = 9, 10$  and 11, regardless of which of the three values of  $\kappa$  is used. Indeed, the prior probability of ten sources specified by the prior with  $\kappa = 10$  is nearly 1.25 million times that of the probability specified by the prior with  $\kappa = 1$ . Despite the difference in the prior probability as a function of  $\kappa$ , the posterior probabilities of  $K = 10$  are quite consistent; the average (across simulations) differs by only about 0.1 (comparing  $\kappa = 1$  with  $\kappa = 10$ , see Figure 2.5). In other words, there is about a 1.25 multiplicative increase in the posterior probability of ten sources when the prior mean is changed from  $\kappa = 1$  to  $\kappa = 10$ . This modest difference in posterior probability is acceptable as it is unlikely that prior information would allocate the truth 1.25 million to one odds.

---

<sup>22</sup>For the purposes of convergence diagnostics, we initialized each chain by randomly choosing between 1 and 20 sources and then deterministically spreading them out around the edge of the image space.



**Figure 2.5:** Average posterior probabilities of plausible values of  $K$  across ten datasets. Left plots show posteriors for the ten-source reality ( $K_{\text{true}} = 10$ ) with prior mean values of  $\kappa = 1, 3, 10$  from top to bottom. Right plots show posteriors for the one-source reality ( $K_{\text{true}} = 1$ ) with  $\kappa = 1, 3, 10$ . In each plot, the 25% and 75% quantiles across the 10 datasets are indicated by the vertical error bars for each value of  $K$ .

There are appreciable differences among the simulated datasets as indicated by the quantiles in Figure 2.5. This is to be expected because the source positions and intensities are chosen randomly. Some of the simulated datasets have two sources very close to each other, making it hard to determine that they are distinct. In some cases, it is possible to separate these very close sources based on the spectral data (using the full model), i.e., if the spectral data appear to come from two *gamma* distributions rather than one. However, in other cases it is difficult to separate such nearby sources, even with the spectral data. Indeed, checks confirmed that datasets with sizeable posterior probability at  $K = 9$  under the full model include overlapping sources that cannot be separated by eye and have similar spectral distributions. Posterior probability at  $K$  values of 11 and above appear because chance clusters of photons are sometimes mistaken for separate sources. The precise location of these ‘ghost’ sources, however, is highly erratic across RJMCMC iterations. There is limited evidence for them in the data and thus wide error bars for their “locations” in the posterior distribution.

Inference is also robust to the choice of  $\kappa$  under the one-source reality ( $K_{\text{true}} = 1$ ). The posterior mode is clearly  $K = 1$  for all three values of  $\kappa$ . Owing to the skewness of the Poisson density, the difference in prior probability of  $K = 1$  across the different  $\kappa$  values is less dramatic than that for  $K = 10$ . When  $\kappa = 1$  the *a priori* probability of  $K = 1$  is around 800 times that when  $\kappa = 10$ . Consequently, the difference in posterior probabilities is also less noticeable. Indeed, the qualitative difference in the posteriors under  $\kappa = 1$  and  $\kappa = 10$  is marginal, see Figure 2.5.

Our key conclusion is that the posterior probability of the true number of sources  $K$  seems insensitive to the prior probability assigned to  $K$ , at least when using the Poisson prior. Consequently, the value of  $\kappa$  only needs to be in the region of the true number of sources in order for the fit for  $K$  to be reasonable. These conclusions match our intuition that knowing the precise PSF statistically constrains the mixture model sufficiently for the data to drive the fitted values of the parameters. Our simulations are representative of typical datasets, but establishing similar conclusions for smaller datasets may require more studies. A dataset could also be larger than those in our simulations, but as  $\kappa$  (and  $K$ ) increases, greater Poisson variance means that the absolute deviation of  $\kappa$  from the true number of sources has progressively less influence on posterior inferences. (Intuitively, it is more reasonable to *a priori* suspect 101 sources when there are 110, than to suspect 1 when there are 10). In our context, prior information typically consists of previous observations, possibly from a different wavelength band. Therefore, it can be assumed that the information is quite reliable and gross prior ‘misspecification’ is unlikely. Clearly, priors other than the Poisson distribution can be considered if a more diffuse prior distribution is desired.

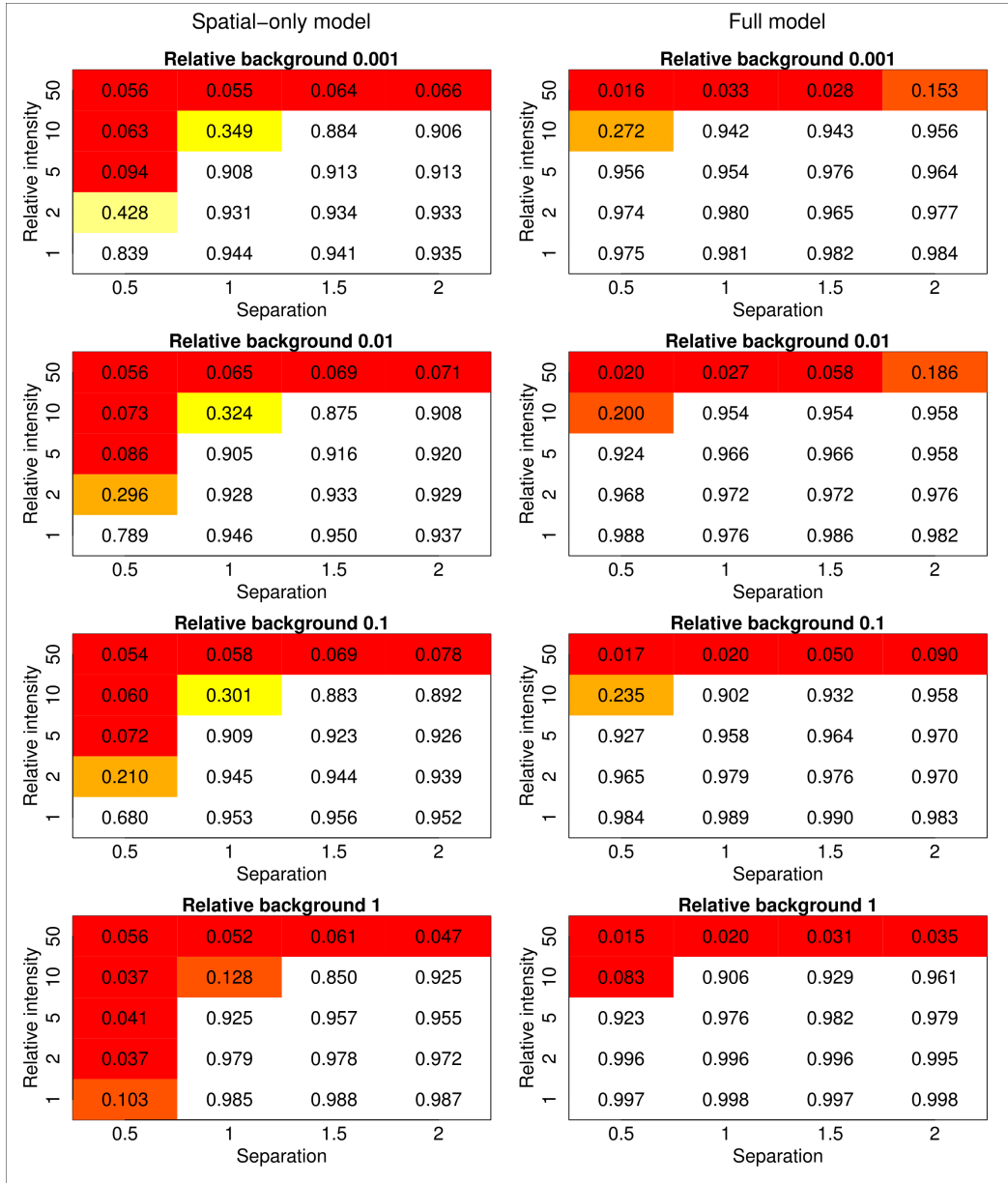
### 2.5.2 Utility of the spectral model

Here we investigate the performance of our model and methods for a range of background intensities, source separations and relative source intensities. We compare the performance of the spatial-only and full models. For simplicity, we simulated data for a two-source ( $K_{\text{true}} = 2$ ) reality. In each simulation, the number

of photons from the background and the number from each source were drawn from Poisson distributions with respective means  $m_0, m_1, m_2$ . We set  $m_2 = 1000$  and  $m_1 = m_2/r$ , for  $r = 1, 2, 5, 10, 50$ . We refer to  $r$  as the relative intensity of the two sources. To set  $m_0$  and quantify the strength of the simulated background in an astronomically meaningful way we define a source region in terms of the PSF. Specifically, we again use the King profile PSF and define the source region as the region with PSF greater than 10% of its maximum. (The King profile density has no finite moments). We next define  $q$  to be the probability that a photon from a source falls within its source region and set the background per source region to be  $m_0 = bqm_2$ , for  $b = 0.001, 0.01, 0.1, 1$ . That is, the mean number of background photons in the faint source region was varied between 1/1000 and 1 times the mean number of photons from the faint source falling in the same region. As we shall discuss and unsurprisingly, the faint source was difficult to locate in datasets that were simulated with  $b = 1$  and less so for those simulated with  $b = 0.001$ . Finally, the separation of the two sources was set to be 0.5, 1, 1.5 or 2 distance units. These separations can be interpreted using the fact that our source regions are approximately circles of radius 1.

Spectral data was also simulated for source and background photons. An aim of this simulation study aims is to investigate how much using the spectral data improves the fitted parameters. Since sources can only be distinguished by their spectra if their spectra are different, we used different spectra for the two simulated sources; specifically we set  $\alpha_1 = 3$ ,  $\gamma_1 = 600$ ,  $\alpha_2 = 6$  and  $\gamma_2 = 1500$ .

In summary, our simulation study consists of a  $5 \times 4 \times 4$  grid of configura-



**Figure 2.6:** Exploring the sensitivity of our algorithm to source separation, relative strengths, and background level. The median posterior probability of  $K = 2$  across the 100 simulations is shown;  $K_{\text{true}} = 2$  in all cases. The results from the spatial-only model (left column) and the full model (right column) are both shown. Red indicates probabilities less than 0.1, and white indicates probabilities greater than 0.5. (Intermediate colors indicate probabilities between 0.1 and 0.5.)

tion settings ( $r = 1, 2, 3, 5, 10, 50$ ;  $b = 0.001, 0.01, 0.1, 1$ ; and source separations of 0.5, 1, 1.5, 2). One hundred datasets were simulated for each of the resulting 80 configurations, and analyzed using first the spatial-only model and then the full model. In particular, for each dataset our algorithm was run for 20,000 RJMCMC iterations, the first 10,000 of which formed the convergence period (or burnin) and were discarded.<sup>23</sup> The median posterior probability of two sources is shown in Figure 2.6 for each of the different simulation settings. The left and the right panels correspond to the spatial-only and full models, respectively. We use the median posterior probability across the 100 simulated datasets because in a few simulations the faint source is unusually bright or unusually faint, which noticeably effects the mean posterior probability of two sources. Nevertheless, summaries based on the mean posterior probability are qualitatively very similar, albeit with slightly more noise. We have organized the results by background intensity because in practical applications background is often well determined.

In images simulated with relative intensity 50 the posterior probability of two sources tends to be low. This is because  $r = 50$  corresponds to a faint source intensity of  $m_2 = 20$ , while the brighter source has intensity  $m_1 = 1000$ . Thus, the faint source is typically not bright enough to be distinguished from noise; its photons can be adequately explained as a random cluster formed of photons from the brighter source or the background. In this case the posterior probability peaks sharply at  $K = 1$ . The spatial-only model is more likely to mistake a cluster of background photons for a faint source and therefore, in the case of  $r = 50$  and

---

<sup>23</sup>This is a relatively small number of RJMCMC iterations, but since our simulated datasets were quite small images each including only two sources, we found it to be sufficient.

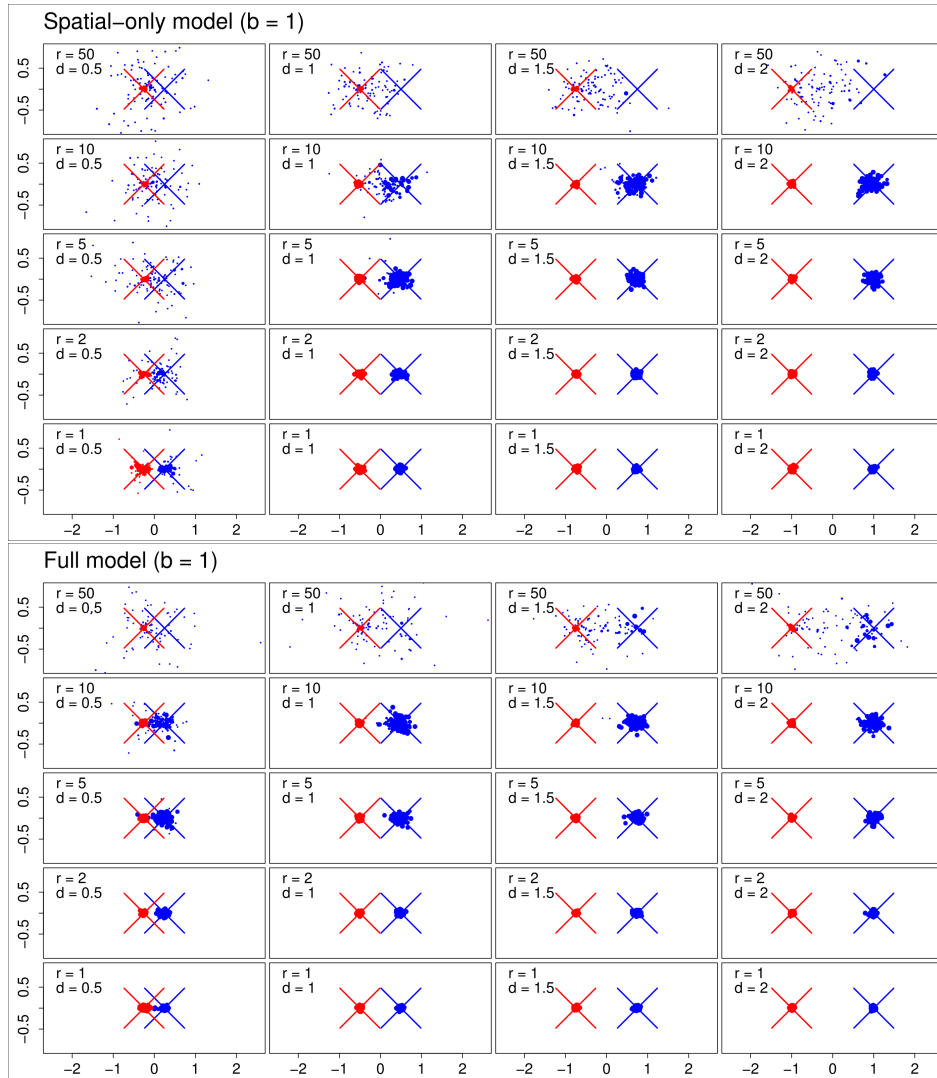


small source separation, typically gives slightly higher posterior probabilities of two sources than the full model (but the probabilities are still very small). For less extreme relative intensities, using the full model increases the posterior probability of two sources. The improvement is particularly noticeable for relative intensities 5 and 10, regardless of the background strength. The spectral distribution of source counts reduces the plausibility that the faint source is just a cluster of photons from the background or the bright source. When both sources are bright and reasonably separated both the spatial-only and full models give high posterior probability at  $K = 2$ .<sup>24</sup>

To fit the source parameters, we fix  $K$  at its posterior mode value and use (2.19). Although the fitted parameters of the bright source are always accurate, those for the faint source may be poor, especially if the posterior mode of  $K$  is at 1 or if “ghost” sources have appreciable posterior probability. The accuracy of the faint source’s fitted parameters essentially follows the pattern seen in Figure 2.6. When the real faint source is very weak or located too close to the bright source, then a fitted second source (when the posterior mode of  $K$  is greater than 1) is likely to be a “ghost” consisting mainly of a cluster of photons from the background or the bright source. In which case, its fitted parameters bear little resemblance

---

<sup>24</sup>One curiosity, present in the left panels of Figure 2.6 (spatial-only model), is that when both of the sources are reasonably bright, greater median posterior probability of two sources is obtained when the background is *stronger*. This phenomenon occurs because, in the presence of strong background, deviations between the PSF and the observed counts are difficult to detect, whereas, with weak background, such deviations may be attributed to spurious additional sources. (Indeed, the posterior probability of  $K = 3$  is typically greater at low background levels than at high background levels). When the full model is used this effect is diminished. The curiosity is not qualitatively important because the bright sources are well identified in all cases. Clearly weaker background is preferred as it improves the chance of detecting (real) faint sources.



**Figure 2.7:** Sensitivity of location determination as a function of source separation, relative strength, and background level. The simulation is the same as that in Figure 2.6. Mean posterior locations of two sources for each of 100 simulations, under the spatial-only model (top 20 plots) and the full model (bottom 20 plots). Red and blue dots give the mean posterior locations for each simulation of the bright and faint sources respectively. The large ‘X’s of corresponding color indicate the true locations. The diameters of the dots are proportional to the posterior probabilities of two sources. The relative background, relative source intensity, and source separation are indicated by  $b$ ,  $r$  and  $d$  respectively.

to those of the true faint source. This is illustrated in Figure 2.7, which shows the mean (conditional on  $K = 2$ ) posterior locations of the two sources for all 100 datasets under each configuration of simulation settings. Crosses indicate the true locations of the sources. The mean posterior locations of the bright source (red dots) are not always visible in the plots because they are often in the middle of the red crosses. The location of the bright source becomes slightly harder to fit as the intensity of the faint source increases. (This is at least partly because the background intensity is proportional to the faint source intensity). The size of the dots indicate the posterior probability of two sources.

The full model again yields more accurate fits. The fitted locations of the faint source (blue dots) center around its true location (blue crosses) for  $r \leq 10$ , even when the source separation is small. For the spatial-only model there is more scatter. Under both models, when  $r = 50$  we can see that many of the fitted faint source locations correspond to spurious clusters of photons surrounding the bright source. As the separation increases some of the fitted faint source locations are halfway between the true locations of the two sources. This occurs when the posterior distribution of the faint source  $x$ -coordinate is bimodal, a spurious cluster of photons and the real faint source both being supported as possible second sources. In an actual analysis this bimodal behaviour would be apparent from inspection of the posterior draws of the source location. For  $r = 50$  and large separation, the full model sometimes accurately fits the faint source location, but the spatial-only model never does. The behavior of the other fitted parameters follows the same pattern illustrated in Figure 2.7 because the fitted

source locations indicate how well photons are allocated to the correct source. This is confirmed by inspecting tables of the mean (or median) squared error of each parameter (not shown).

The number of Monte Carlo samples used in estimating the mean posterior locations (conditional on  $K = 2$ ) is determined by the posterior probability of two sources, and thus is indicated by the size of the dots. Very small dots may have non-negligible Monte Carlo error i.e. the true posterior mean location (conditional on  $K = 2$ ) may be somewhat inaccurately approximated. This is because applying (2.19) for each parameter does not accurately compute the mean of  $p(\Theta_K|K, x, y, E)$  for values of  $K$  that have low posterior probability.<sup>25</sup> However, in practice, when the number of sources is unknown, it makes sense to only consider values of  $K$  with relatively high posterior probability. Furthermore, one typically checks the level of Monte Carlo error for the values of  $K$  of interest, by running multiple chains. Large variation in the parameter estimates across the chains indicates high Monte Carlo error. In which case, one should run the chains longer in order to obtain a larger Monte Carlo sample.

## 2.6 Application I: XMM dataset

We now apply the spatial-only and full models to an XMM observation (obs\_id 0151450101) of the apparent visual binary FK Aqr and FL Aqr. The data con-

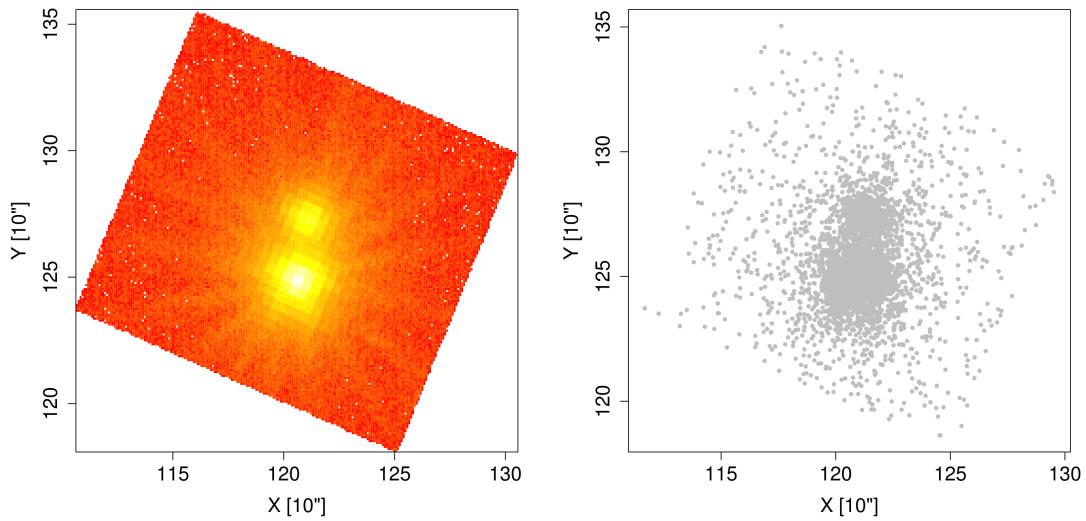
---

<sup>25</sup>We could instead fix  $K = 2$  and run a standard MCMC algorithm to obtain a large enough posterior sample to accurately fit the mean posterior locations. We do not pursue this strategy because the fitted parameters that are conditional on unlikely values of  $K$  are of little practical use.

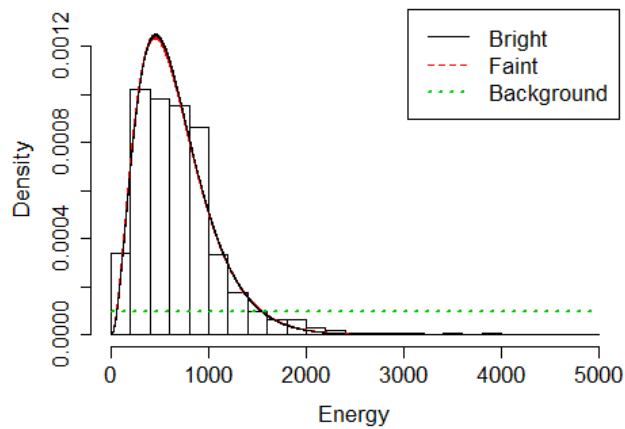
sist of the spatial and spectral information of around 540,000 photons detected during a 47ks exposure. The spatial data is displayed in Figure 2.8 as both an image (left) and a scatter plot (right), and the spectrum is plotted in Figure 2.9. The moderate overlap of the sources and high counts make this a good test of our model. In particular, we expect that the spatial-only model and full model analyses to be similar (for the spatial parameters) because of the large amount of spatial information. Furthermore, since the data clearly indicate two sources, we can concentrate on verifying that our model yields sensible posterior inference using standard MCMC. (This gives draws from the joint posterior for a fixed number of sources and therefore results in inference that is simpler to interpret than inference resulting from RJMCMC.) Use of the more complicated RJMCMC analysis is reserved for the *Chandra* dataset in Section 2.7 because there is non-negligible uncertainty in  $K$  for that dataset.

In the image shown on the left of Figure 2.8 the sources seem to have faint ‘spokes’. Approaches for modeling these features are suggested in [Read et al. \(2010\)](#) and [Read and Saxton \(2012\)](#), but we use the unaltered King profile PSF for simplicity. As mentioned in Section 2.2.1, the spatial data are binned when recorded on the observatory LCD screen. However, the bins are small in comparison to the XMM PSF so our use of a model that treats the data as unbinned is reasonable. (See Section 2.8 for further discussion.)

For the spatial-only model and the full model, ten MCMC chains (with  $K$  fixed at 2) were run for 20,000 MCMC iterations, the first 10,000 of which formed



**Figure 2.8:** Visual binary FK and FL Aqr observed with XMM-Newton (FK is the brighter source at bottom). The XMM obs\_id is 0151450101. Shown is a counts image with  $10''$  bins and arbitrary origin (left), and a scatter plot of a subset of 6,000 events over a 5ks subexposure (right).



**Figure 2.9:** A histogram of the spectral data in the XMM observation of FK Aqr and FL Aqr. Plotted are 1,000 spectra for the bright (solid black lines) and faint (dashed red lines) sources, each corresponds to a posterior sample of the spectral parameters. (The posterior variance is small on this scale.) The background spectra is shown by the dotted green line.

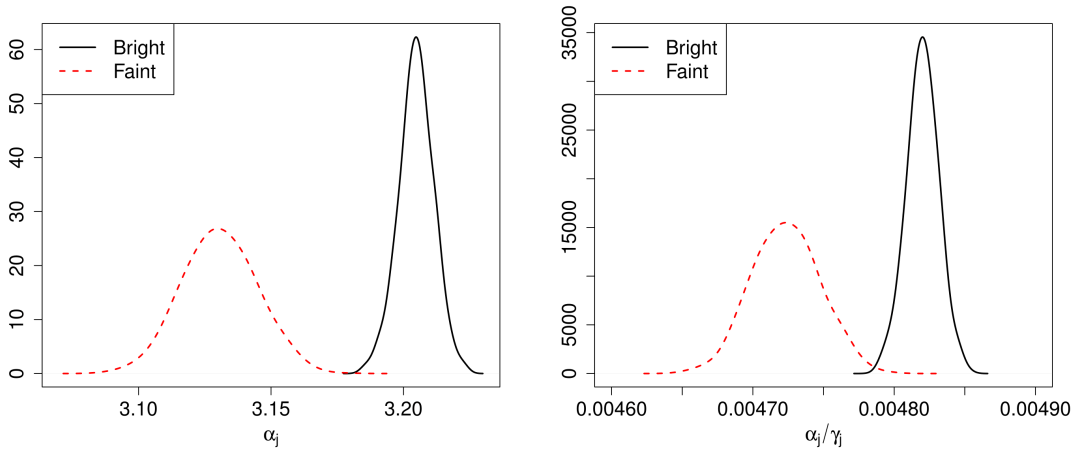
**Table 2.4:** Posterior means under the spatial-only model and the full model. The parenthetic intervals are  $1\sigma$  error bars computed using 16% and 84% posterior quantiles.

	Spatial-only model		Full model	
$\mu_{1x}$	120.974	(120.973,120.975)	120.973	(120.973,120.974)
$\mu_{1y}$	124.873	(124.873,124.874)	124.873	(124.872,124.874)
$\mu_{2x}$	121.396	(121.394,121.398)	121.397	(121.395,121.399)
$\mu_{2y}$	127.319	(127.317,127.321)	127.326	(127.324,127.328)
$w_1$	0.717	(0.716,0.718)	0.732	(0.731,0.732)
$w_2$	0.182	(0.181,0.182)	0.189	(0.189,0.190)
$w_0$	0.102	(0.101,0.102)	0.079	(0.079,0.079)
$\gamma_1$	–	–	664.86	(664.43,665.30)
$\gamma_2$	–	–	662.78	(661.78,663.87)
$\alpha_1$	–	–	3.205	(3.199,3.211)
$\alpha_2$	–	–	3.131	(3.118,3.144)

the convergence period (or burnin) and were discarded.<sup>26</sup> The large amount of data means that the source locations are precisely fit by both models, as can be seen in Table 2.4. However, the posterior mean of the relative intensity of the background is about 20% lower for the full model. This is presumably due to a greater ability to separate source and background counts with the additional information given by the spectral data. In particular, photons from the sources can be found across the entire image so there is a tendency to over-estimate the background intensity without some non-spatial way of distinguishing its photons from those of the sources.

Until now, it has not been possible to distinguish the spectral distributions of these two sources. Conventional fitting of the spectra extracted from non-overlapping source regions give statistically indistinguishable results, with identical column density  $N_{\text{H}} \approx 1.0 - 1.6$  ( $10^{20}$   $\text{cm}^{-2}$ ), double temperature components  $kT_1 \approx 0.25 - 0.26$  (keV),  $kT_2 \approx 0.78 - 0.82$  (keV), and metallicities  $Z \approx 0.12 - 0.14$ .

<sup>26</sup>Note that, since we used standard MCMC and there are only two bright sources, the number of MCMC iterations until convergence was relatively small.



**Figure 2.10:** Posterior distributions of the parameters of the *gamma* distributions used to model the spectra of FK Aqr and FL Aqr. The posterior distributions of the shape and rate parameters are shown in the left and right panels, respectively.

This remarkable coincidence could be attributed to strong contamination of FL Aqr by photons from FK Aqr. Our algorithm, which eliminates such contamination, can answer the question of how similar the two sources are. Of course, a comparison of the source spectra shapes is only possible using the full model. Figure 2.9 shows 1,000 spectra sampled from the posterior distribution<sup>27</sup> for the bright (black solid lines) and faint (red dashed lines) sources; for each source, all 1,000 spectra are very similar and so appear as a single curve. We observe that the bright source spectra are very similar to the faint source spectra, which is consistent with the difficulty in distinguishing the spectral distributions of the two sources in previous analyses.

Although the overall shapes of the two spectral are similar (Figure 2.9), we can

<sup>27</sup>To reduce correlation, every 10<sup>th</sup> sample of the original 10,000 stored MCMC samples of the spectral parameters was used.



distinguish them by examining the parameters of their underlying *gamma* distribution. Figure 2.10 plots the posterior distributions of these parameters for the two sources and shows that they clearly differ. We have plotted the shape and rate parameters, because the shape and variance differ more than the mean. The posterior distributions in Figure 2.10 indicate that there is very little uncertainty in the spectral parameters; the intervals in Table 2.4 convey a similar message. This precision is obtained because of the large amount of data combined with the fact that our method properly accounts for uncertainty in photon origins and jointly fits spectral and spatial parameters. Although our analysis is only physically accurate to the extent that the source spectra can reasonably be modeled with *gamma* distributions, it nevertheless provides evidence that the spectra do differ in some way. More detailed conclusions would be possible with a physics-based spectral model that accounts for emission lines and other spectral features. A possible extension of this work is to replace the *gamma* spectral model with a more complete model. A computationally less intensive approach is described in Section 2.7.2.

## 2.7 Application II: Chandra dataset

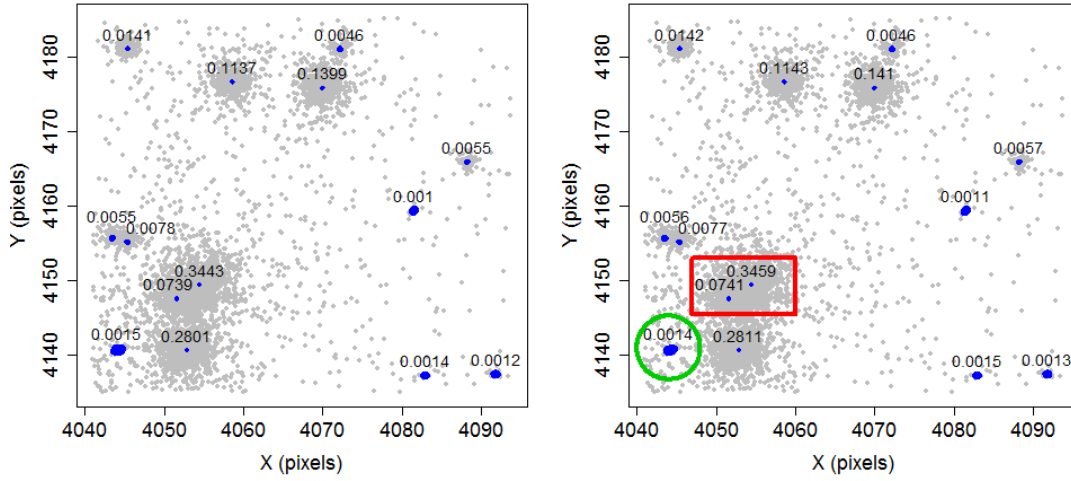
We analyze a *Chandra* observation of the Orion Nebula Cluster using the spatial-only model and the extended full model given in (2.11). The extended full model is used because the full model is not sufficiently flexible to capture the shape of the source spectra, as explained in Section 2.2.4. The specific dataset we analyze

is a subset of ObsID 1522 that omits the central source, a region where the PSF is distorted due to strong pile-up (Figure 2.11). The data include events that occurred within the first 20ks of the observation, of which there are  $\approx 14,000$ .

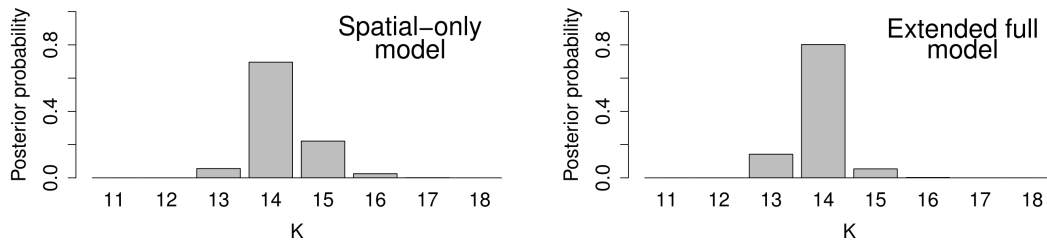
### 2.7.1 Spatial-only and extended full model analyses

For both models, ten RJMCMC chains were run for 150,000 RJMCMC iterations, the first 100,000 of which formed the convergence period (or burnin) and were discarded. The posterior distribution of the number of sources, under the spatial and extended full models, is displayed in Figure 2.12. The mode of both posteriors is at 14. However, the spatial-only model shows slightly more uncertainty, and some support for 15 sources.

As mentioned in Section 2.4,  $K$  determines the number and meaning of the other model parameters and therefore we must condition on a value of  $K$  to draw meaningful inferences for them from the RJMCMC output. Figure 2.11 shows 90% posterior credible regions (blue) for the locations of the sources under the two models, given  $K = 14$ . Each credible region shows an area which has 0.9 posterior probability (given  $K = 14$ ) of containing the location of the relevant source, i.e., an integral of the posterior distribution of the source location (given  $K = 14$ ) over this area would evaluate to 0.9. The credible regions look to be similar under the two models. The estimated relative intensities also appear in Figure 2.11 and are also similar, but are slightly lower under the spatial-only model for most sources. This is due to a higher estimate of the relative background intensity under the



**Figure 2.11:** Chandra observation of a crowded field near the center of the Orion Nebula Cluster. This field is approximately  $25'' \times 25''$  in size, and is centered at  $(RA, Dec) = (5:35:15.4, -05:23:04.68)$ . Shown in blue are approximate 90% posterior credible regions for source locations, under the spatial-only model (left), and the extended full model (right). The figures next to the regions indicate the estimated relative intensities. The credible region of the source with the largest location uncertainty is circled in green (right panel). The red rectangular box encloses two overlapping sources (right panel) for which we carry out a detailed follow-up spectral analysis (Section 2.7.2).



**Figure 2.12:** Number of sources detected in the analysis of the Chandra observation in Figure 2.11. Posterior of  $K$  based on the spatial-only model (left) and the extended full model (right).

**Table 2.5:** Extended full model fit for the Chandra observation in Figure 2.11. Posterior mean locations and relative intensities (as percentages), with 68% intervals indicated.

COUP #	$\mu_{jx}$	$\mu_{jy}$	Relative intensity (%)
732	4054.42 (4054.41,4054.43)	4149.45 (4149.44,4149.46)	34.59 (34.16,35.03)
745	4052.83 (4052.81,4052.84)	4140.67 (4140.66,4140.68)	28.11 (27.71,28.51)
689	4069.93 (4069.91,4069.94)	4175.93 (4175.91,4175.94)	14.10 (13.79,14.40)
724	4058.57 (4058.56,4058.59)	4176.73 (4176.71,4176.74)	11.43 (11.16,11.71)
744	4051.53 (4051.50,4051.55)	4147.57 (4147.55,4147.60)	7.41 (7.14,7.68)
765	4045.40 (4045.35,4045.46)	4181.20 (4181.15,4181.25)	1.42 (1.32,1.53)
649	4088.16 (4088.08,4088.24)	4165.95 (4165.87,4166.03)	0.57 (0.50,0.63)
766	4045.36 (4045.27,4045.45)	4155.18 (4155.10,4155.25)	0.77 (0.68,0.87)
788	4043.48 (4043.36,4043.61)	4155.74 (4155.64,4155.84)	0.56 (0.47,0.64)
682	4072.11 (4072.01,4072.21)	4181.12 (4181.03,4181.22)	0.46 (0.39,0.52)
640	4091.73 (4091.53,4091.92)	4137.42 (4137.26,4137.59)	0.13 (0.10,0.16)
664	4081.43 (4081.22,4081.63)	4159.41 (4159.21,4159.61)	0.11 (0.08,0.14)
665	4082.84 (4082.67,4083.02)	4137.28 (4137.14,4137.43)	0.15 (0.12,0.19)
779	4044.39 (4043.86,4044.60)	4140.72 (4140.43,4140.90)	0.14 (0.09,0.18)
Background	–	–	0.06 (0.01,0.10)

spatial-only model (0.0053 versus 0.0006 under the extended full model<sup>28</sup>). Table 2.5 gives the the posterior mean fit of the source locations and relative intensities under the extended full model for  $K = 14$ . The detected sources are also matched to the source catalog from the Chandra Orion Ultradeep Project (COUP; [Getman et al. 2005](#)).

Other observations of Orion suggest that the source circled (in green) in the

<sup>28</sup>The background is likely inaccurately estimated by both models because the King profile PSF that we use is an approximation to the *Chandra* PSF; the latter is more concentrated at its center. Thus, in our analysis, too many photons are allocated to the wings of the sources, deflating the background. That our analysis has still found genuine sources illustrates that it is not too sensitive to the PSF, at least in the case of specifying overly heavy wings. If instead the raytraced PSF (ChaRT: <http://cxc.cfa.harvard.edu/chart/>) is used, then the estimate of the background is higher because this PSF has lighter wings than the King profile. The lighter wings also lead to the detection of four additional faint sources: one has an optical counterpart, one does not, and two cannot be confirmed optically because they are close to a bright source. Further investigation of these sources and modeling possible variations in the PSF are topics for future work. For example, temporal information can potentially be used as a diagnostic to assess whether any of the detected weak sources are in fact due to fluctuations in the PSFs of the bright sources.

right panel of Figure 2.11 is a genuine source. Its location is more uncertain than other sources because it is more difficult to detect. Indeed, with an estimated intensity between 13 and 25 counts, this source is at the edge of detectability of local detection methods, particularly since the estimate of the local background in such methods would be high due to contamination from nearby bright sources. Thus, we expect that more basic approaches would either have failed to find this source, or would only find it by rendering their detection threshold to a point where spurious detections became problematic. Indeed, the reason the spatial-only model gives non-negligible weight to 15 sources (see Figure 2.12) is that it tends to split sources into two. The problem is that a single empirical PSF may exhibit chance variations that appear to be evidence for multiple PSFs. The spatial-only model also mistakes clusters of background photons for sources. The locations and spectra of these spurious sources show considerable posterior variability. Although any particular instance has low probability, there are multiple instances that together create erroneous support for an additional source. The main advantage of using the spectral information, in this example, is that it mitigates these issues, leading to a greater certainty that there are really 14 sources. Additionally, under the extended full model, the standard deviations of the parameters are almost invariably slightly smaller.

## 2.7.2 Spectral analysis of the disentangled sources

The extended full model only captures the basic shape of the source spectra and we now illustrate how detailed follow-up spectral analysis can incorporate proba-

bilistic event allocations. We perform this analysis for the two overlapping sources that are enclosed in the red box in the right panel of Figure 2.11 (COUP sources #732 and #744 (Getman et al. 2005)). Their estimated relative intensities are 0.3459 and 0.0741 under the extended full model. This is a good example to test the probabilistic event allocations, since the sources are close together (separation  $\approx 1.7''$ ), each have sufficient counts for a useful spectral fit ( $\approx 4350$  and  $\approx 910$  counts between  $0.5 - 7$  keV for the bright and faint sources, respectively), and one source is substantially weaker than the other.

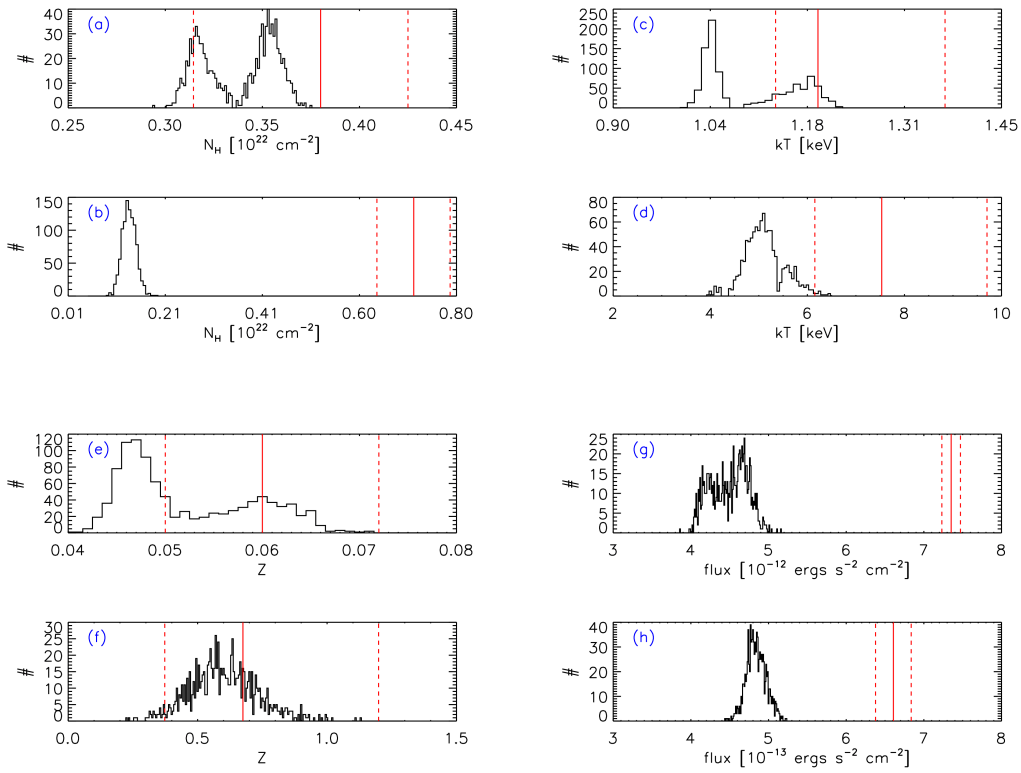
As described in Section 2.2.3,  $s_i$  indicates the source (or background) number associated with photon  $i$ . These are unknown parameters (or latent variables) that are updated at each iteration of the RJMCMC sampler. The variability in  $s_i$  indicates the uncertainty in the source of photon  $i$  (due to the PSF and uncertainty in the source parameters). We can account for this uncertainty by conducting many spectral analyses, each according to a sampled photon allocation (i.e., sampled values of  $s_i$ ), and combining the results. We focus on photons with spatial location in the red box in Figure 2.11 (right panel) and to values of  $s_i$  sampled conditional on  $K = 14$ . Since we are only interested in COUP sources #732 and #744 we ignore any photons that are attributed to one of the other sources (in a given allocation). (The photons in the red box in Figure 2.11 are attributed to one of the other sources only rarely).

Based on the photon allocations, we construct a sample of 1000 simulated spectral datasets for both sources, constructed from photon allocations based on every  $10^{\text{th}}$  iteration of the RJMCMC algorithm that sets  $K = 14$  (up to the  $10,000^{\text{th}}$

RJMCMC iteration that sets  $K = 14$ ). The variability in the source counts across the 1,000 iterations is  $\pm 17$  for both the bright and faint sources. The specific photons that are allocated to each source also varies, even when the total source counts do not. Each individual spectrum is fit with an absorbed single temperature thermal model (`xsphabs*xsapec` in CIAO/*Sherpa* v4.6) fitting the absorption column ( $N_{\text{H}}$ ), temperature ( $kT$ ), metallicity ( $Z$ ), and normalization. A pile-up correction is needed for all spectra for the bright source since the measured count rate of 0.7 counts frame<sup>-1</sup> is higher than the threshold at which pile-up becomes significant ( $\approx 0.3$  counts frame<sup>-1</sup>). We use the `jdpileup` model in *Sherpa*, fitting the grade migration parameter  $\alpha$  and the pile-up strength parameter  $f$  (Davis 2001). We call the entire collection of spectral fits the disentangled analysis.

For comparison, we also carry out a spectral analysis of the sources based on a naïve allocation of photons that collects events from within 1'' of the fitted location of each source and assumes that there is no contamination from the other source. The only difference in the spectral model for the naïve and disentangled analyses is in how the effective areas are defined. In the case of the naïve analysis, a correction is made post-facto to the normalization based on how much of the source is expected to be included within the 1'' source photons extraction radius. In the disentangled analysis, the assumed extraction radius for the spectra with allocated events is set to be 2.5'' and the subsequent correction is negligible.

The results of the spectral fits to the disentangled spectra are shown as histograms of best-fit values for  $N_{\text{H}}$ ,  $kT$ ,  $Z$ , and model flux computed for each of the 1000 spectra, see Figure 2.13. In several cases, a bimodal distribution is apparent.



**Figure 2.13:** Detailed spectral analysis of overlapping COUP sources #732 and #744. Best-fit values of absorption column ((a), (b)), temperature ((c), (d)), metallicity ((e), (f)), and flux ((g), (h)) for the disentangled analysis, for each of 1000 allocations of the photons are shown as histograms. Panels (a), (c), (e), and (g) correspond to the bright source and panels (b), (d), (f), and (h) correspond to the fainter source. The naïve analysis best-fit values and their 68% intervals are shown by the solid and dashed red vertical lines, respectively. The width of the histograms only account for uncertainty due to the allocation of photons, and not additional statistical error, which is well described by the intervals shown for the naïve analysis.



This suggests that a multi-temperature component spectrum would be a better fit. The separation of the modes, however, is generally too small to be picked up by typical multi-temperature model fits. Not shown are the pile-up parameters for the bright source, which are consistent between the naïve and disentangled analyses ( $(\alpha, f) = (0.6, 0.93)$  for naïve, and  $(0.53, 0.89)$  for the disentangled spectra), though the former indicates that the pile-up strength is slightly higher. This is to be expected, since the naïve analysis is carried out for photons in the core of the PSF, where naturally pile-up is most significant. The disentangled spectra include photons from the wings, thus reducing the strength of pile-up effects and decreasing the correction needed to the source flux by about 60%.

The spread in the histograms in Figure 2.13 indicates the uncertainty in the best-fit values due to uncertainty in the allocation of photons. The best-fit values from the naïve calculation are shown as solid red vertical lines. The dashed red vertical lines give 68% intervals indicating the statistical errors, due to randomness in the photons emitted and detected, under the naïve analysis. These statistical errors do not account for uncertainty in the photon allocations. The histograms, on the other hand, represent only errors due to uncertainty in the photon allocations, but do not account for statistical errors (due to randomness in photon emission and detection). Because the two sources of error are independent, and because we expect the statistical errors for the disentangled analyses to be similar to those for the naïve analysis, the total errors could be represented by a perturbation of the histograms with  $\sigma$  equal to the statistical errors from the naïve analysis. For these data, with the exception of flux (panels (g) and (h)

of Figure 2.13), the statistical errors dominate the errors due to uncertainty in the photon allocation. Despite this, the disentangled analysis provides reasonable evidence that the absorption column of the faint source (panels (b) of Figure 2.13) and the flux of the two sources (panels (g) and (h) of Figure 2.13) are different from the best-fit values under the naïve analysis.

The variability of the true parameters around each of the best fit values recorded in the histograms is expected to be similar to that indicated for the naïve fit. However, we did not calculate these uncertainties because of the large computational cost. For these data, with the exception of flux (panels (g) and (h) of Figure 2.13), the variability in the true spectral parameters around the best fit values is likely larger than the uncertainty in the best fit values (due to the uncertainty in the allocation of photons). Despite this, the disentangled analysis provides reasonable evidence that the absorption column of the faint source (panel (b) of Figure 2.13) and the flux of the two sources (panels (g) and (h) of Figure 2.13) are different to the naïve analysis best-fit values.

Overall, the naïve analysis best-fit values for the fainter source are in greater disagreement with the disentangled analysis than those for the brighter source. This is to be expected, since in the naïve analysis, the contamination of the fainter source by the brighter source is larger. Our algorithm effectively removes this contamination. This causes the spectral fit parameter values to change and the measured source flux of the fainter source to decrease. In summary, the observed changes to the spectral model parameters are as would be expected when contamination is reduced and the data quality is improved.

## 2.8 Summary

We have developed a Bayesian statistical method that models spatial and spectral information from overlapping sources and the background, and jointly estimates all individual source parameters. The key contributions of our approach are the use of spectral information to improve spatial separation, coherent quantification of uncertainty, including that of the number of sources, and the probabilistic assignment of photons to the different sources. Our simulation studies show that using spectral information improves the detection of both faint and closely overlapping sources and increases the accuracy with which source parameters are inferred.

We have analyzed data from two sets of overlapping sources observed with XMM and *Chandra*. Traditional analysis of XMM observations of FK and FL Aqr, thought to be a visual binary, show that their spectra are not distinguishable. Our analysis confirms that the spectra are indeed similar, but nonetheless shows that they are separable. We have also carried out detailed spectral analysis on disentangled photons from a pair of close sources from near the center of the Orion Nebula Cluster observed with *Chandra*. We find that the spectral parameters change significantly after contamination is removed.

The data we have considered consists of event-level observations. In the more usual case of spatially binned data, the PSF could be updated to take account of the binning. If the spatial pixels are larger, the importance of spectral data is greater, because it is harder to spatially distinguish sources from each other and the background. Clearly however, unbinned data is preferred when available,

and our method has the ability to use all the information in such data. Similar comments apply when the spectral data are grouped.

As with other detection procedures, an important question is how to combine information from multiple observations. Since our approach gives the posterior distribution of all the parameters, this can be used as the prior distribution in subsequent analyses. Thus, under the Bayesian framework it is straightforward to analyze the available observations sequentially, which is convenient in that different PSFs, for example, can be used for each analysis. This is critical if the observations are recorded by different observatories.

Another advantage of the Bayesian framework is that more complex models can straightforwardly be built in. For example, using a location or spectral dependent PSF would require only minimal changes to the method and code. Another extension is to include the different temporal signatures of overlapping sources to further separate them. Future work will focus on these and related issues as well as computational scalability.

# 3

## Likelihood methods for Monte Carlo estimation

### 3.1 Introduction

The normalizing constant of a density, or more generally an arbitrary function  $q$ , is defined as

$$c = \int_{\Omega} q(x) d\mu(x), \quad (3.1)$$

where  $\mu$  is the baseline measure and  $\Omega$  is the support of  $q$ . In statistics, important quantities such as Bayes factors and observed data likelihoods can often be computed by calculating appropriate normalizing constants, e.g., the observed data likelihood is the normalizing constant of the complete data likelihood. Physicists are also routinely interested in normalizing constants, such as partition functions, see for example [Bennett \(1976\)](#) and [Voter \(1985\)](#). However, it is common for normalizing constants to be analytically intractable, and even numerical integration is often infeasible for computing (3.1), especially in high dimensional settings. To tackle this difficulty, a wealth of Monte Carlo methods have been developed to estimate normalizing constants, see for example the many approaches reviewed and proposed in [Liu \(2008\)](#). Broadly, Monte Carlo methods consist of drawing random samples and using them in approximating a quantity of interest, e.g., researchers often use Markov chain Monte Carlo (MCMC) algorithms to draw random but correlated samples from a posterior distribution and then average these samples to estimate the posterior mean. The literature on Monte Carlo approaches has mainly focused on how to draw useful samples in an efficient way, and the second step of constructing an estimate based on the samples has received relatively little attention. Indeed, in practice, the final estimator is often a default moment estimator which is typically not known to be globally optimal in any sense.

To address this gap in the literature, [Kong et al. \(2003\)](#) proposed a maximum likelihood approach for deriving Monte Carlo estimators of normalizing constants by treating the baseline measure  $\mu$  as the unknown and “estimating” it to produce an optimal discrete approximation of integrals such as (3.1). Their method

shows that some familiar Monte Carlo approaches such as importance sampling and bridge sampling (Meng and Wong 1996) are in fact optimal. It also provides a convenient framework for trading off computational efficiency with statistical efficiency through the selective use of our knowledge of the true measure. In particular, by considering *sub-models* that restrict our attention to a certain class of measures, we can obtain more precise estimates with the same number of samples, but still avoid the intractable integral (3.1) that we recover by blindly including all information we have about the measure. One appealing type of sub-model specifies the measure to be invariant to a finite group of transformations, such as particular reflections or rotations. For example, by reflecting about the origin we utilize  $-x$  as well as the sample  $x$  in computing an approximation to (3.1). Thus, the gain in precision comes from increasing the number of points at which the density  $q$  is evaluated. The class of group invariant sub-models is particularly powerful because the baseline measure is nearly always the Lebesgue measure or the counting measure so there are many symmetries that can be exploited. The best sub-model to consider depends on the densities we have samples from and the integrals of interest, but group invariant sub-models are useful in almost any problem. Other options, such as parametrization of the measure, are much less generally applicable in practice.

Two constraints on the invariance groups are that they must be finite and their transformations must leave normalizing constants unchanged, or changed in known ways. The finite group constraint rules out translations and many other useful transformations. This thesis greatly expands the collection of transforma-

tions that the group invariant sub-model framework may exploit by introducing a simple augmentation of the Monte Carlo samples that allows any one-to-one transformation to be used, provided we can compute the transformation Jacobian. Important additions to the group invariant sub-model framework under this extension include the “warp transformations” that [Meng and Schilling \(2002\)](#) suggested and demonstrated to be very effective at increasing the precision of bridge sampling estimators for ratios of normalizing constants, e.g., translations and scalings. As in the bridge sampling context, our main focus will be on estimating ratios of normalizing constants, but the approach is general since we can estimate a ratio in which one of the normalizing constants is known.

The group invariant sub-models we propose *modify* the underlying Monte Carlo samples, as opposed to using them as building blocks to generate more points, which is the idea in the original group invariant sub-model formulation, e.g., in the simple reflection example above,  $x$  is used to generate  $-x$ . Thus, in our extension, estimation precision is gained by using symmetries of the measure to facilitate density evaluations that are more useful, as opposed to using the symmetries to introduce additional density evaluations.

We also explore two methods for optimizing the parameters of an invariance group, such as the line of reflection or the scaling parameter. The first optimization method follows the decision theoretic approach proposed by [Meng and Schilling \(2002\)](#) for choosing warp transformations, and the second method is an approximate likelihood approach. Neither method is entirely satisfactory and the search for an approach that falls completely under the likelihood framework and



gives optimal joint estimation of the estimands and the invariance group parameters is the topic of ongoing work. The difficulties in achieving this are closely linked with the discussion in [Vardi \(1985\)](#) regarding the existence and uniqueness of a non-parametric maximum likelihood estimate of a cumulative distribution function (CDF) based on samples drawn from multiple weighted versions of the CDF, i.e., biased samples. The identification of this challenge underlying the likelihood method proposed by [Kong et al. \(2003\)](#) and the connection to [Vardi \(1985\)](#) is the second main contribution of this chapter since the ultimate success of the overall approach is likely to depend on resolving this issue. In the meantime, despite the challenges in identifying an optimal procedure, the methods we suggest for choosing invariance group parameters are nonetheless better than more basic approaches, such as matching density modes to estimate a translation parameter.

Section 3.2 reviews the maximum likelihood framework for Monte Carlo estimation and group invariance sub-models introduced in [Kong et al. \(2003\)](#) and further developed in [Kong et al. \(2006\)](#). Section 3.3 extends the group invariance sub-model framework to include the warp transformations of [Meng and Schilling \(2002\)](#). Section 3.4 discusses two methods for choosing invariance group parameters, their limitations, and further exploratory ideas to investigate in the search for a fully optimal approach.

## 3.2 Likelihood methods for optimizing Monte Carlo integration

It is often infeasible to compute normalizing constants analytically and so approximations are routinely required. The simplest approach is to sum rectangular areas but this is extremely inefficient in higher dimensions so researchers turn to Monte Carlo techniques. Monte Carlo methods are themselves wide ranging in their efficiency, and a naive application using uniform sampling can similarly run into problems in high dimensions; many of the Monte Carlo samples may be in unimportant regions and the sampling itself can be challenging. To overcome these difficulties, researchers often draw samples from more carefully chosen distributions using methods such as rejection sampling or Monte Carlo Markov chain (MCMC) algorithms. In addition to the question of which distribution(s) should be simulated from and how to do this, there is also the question of what should be done with the samples obtained, which is the focus of this chapter. We do not further discuss exactly how samples are obtained but assume that they are expensive and therefore it is of interest to optimize the efficiency of our estimation procedure given the set of samples. For simplicity, we assume that the samples are independent, which is reasonable since approximately independent samples can often be obtained by, for example, thinning MCMC samples. Even when approximate independence is not achievable, methods such as bridge sampling work with a few simple modifications provided the samples are not too strongly correlated, see [Meng and Wong \(1996\)](#). In any case, understanding the independent

samples case is an essential step to knowing how to proceed in the more complex case of dependent samples.

Given  $n$  samples, the Monte Carlo literature often suggests using moment estimators to estimate normalizing constants or their ratios. For example, given the samples  $x_1, \dots, x_n \sim p_2$ , the well known importance sampling estimator of the normalizing constant  $c_1$  in (3.1) is

$$\hat{c}_1 = \frac{1}{n} \sum_{i=1}^n \frac{q_1(x_i)}{p_2(x_i)}, \quad (3.2)$$

where  $p_s = q_s/c_s$  is a probability density,  $c_s$  is a normalizing constant, and  $q_s$  is the unnormalized density, for  $s = 1, 2$ . We will use  $\Omega_s$  to denote the support of  $p_s$ , for  $s = 1, 2$ , but if the support of all densities under consideration is the same we will simply write  $\Omega$ .

The importance sampling estimate (3.2) is not vastly different from the basic numerical integration method of summing rectangular areas (or volumes) that are of equal width and whose midpoints and heights are  $m_i$  and  $q_1(m_i)$ , respectively, for  $i = 1, \dots, n$ . Indeed, if we draw from a fine bin approximation to  $p_2$  and use the midpoints of the sampled bins as our samples  $x_i = m_i$ , for  $i = 1, \dots, n$ , then (3.2) is a random summation of rectangles with the heights now chosen to be  $\frac{q_1(x_i)}{p_2(x_i)}$ , for  $i = 1, \dots, n$ . The adjusted heights correct for the fact that our samples are concentrated in regions of high density of  $p_2$  rather than covering the whole support of  $p_1$  evenly. (Alternatively, we may view  $1/p_2(x_i)$  as playing an analogous role to the *widths* of rectangles in the basic numerical integration

method.) Provided that  $\Omega_1 \subset \Omega_2$ , the importance sampling estimator (3.2) has the benefit of being unbiased for  $c_1$  without requiring the large computational cost that is incurred by using many rectangles in the basic numerical integration method. However, the variance of (3.2) may be large, particularly if  $p_1$  and  $p_2$  are not similar, and therefore a lot of care may be required in choosing  $p_2$ .

From a fundamental perspective, the appearance of  $1/p_2(x_i)$  in (3.2) is curious because the constant  $c_1$  is unrelated to  $p_2$ . Kong et al. (2003) point out that it is the common baseline measure  $\mu$  of the densities  $p_1$  and  $p_2$  which connects  $c_1$  and  $p_2$ , and that the weight  $(np_2(x_i))^{-1}$  is in fact an *estimate* of  $\mu(x_i)$ , for  $i = 1, \dots, n$ . Furthermore, by maximizing the likelihood, we can show that the approximation  $\hat{\mu}(x_i) = (np_2(x_i))^{-1}$  in (3.2) is optimal. The correct likelihood is

$$L(\mu; x) = \prod_{i=1}^n p_2(x_i)\mu(x_i), \quad (3.3)$$

with the important constraint

$$\int_{\Omega} p_2(x)d\mu(x) = 1. \quad (3.4)$$

Here and elsewhere  $\mu(x_i)$  denotes the measure of the set  $\{x_i\}$ , for  $i = 1, \dots, n$ . Since the baseline measure  $\mu$  is the only unknown quantity, the likelihood is clearly maximized by  $\hat{\mu}(x_i) = (np_2(x_i))^{-1}$  as required. Thus, by invariance of the

maximum likelihood estimate (MLE), the MLE of  $c_1$  is given by

$$\hat{c}_1 = \int_{\Omega} q_1(x) d\hat{\mu}(x) = \frac{1}{n} \sum_{i=1}^n \frac{q_1(x_i)}{p_2(x_i)}, \quad (3.5)$$

which is exactly (3.2). The power of this likelihood formulation is two-fold: (i) it allows us to check the (asymptotic) optimality of (3.2) and other existing estimators and thus conserve research time and effort, and (ii) it provides a principled method for finding optimal estimators in more complex situations where it is difficult to construct good estimators.

An important caveat to keep in mind is that estimating the baseline measure  $\mu$  on the whole real line is neither a tractable nor a useful goal. Our real goal is to obtain an approximation to  $\mu$  in regions where the integrands of interest are large. In order to have some chance of success, we must assume that the integrands and sampling densities have sufficient smoothness and that the measure does as well (but here we assume the baseline measure to be the Lebesgue measure and therefore smoothness constraints are not a concern). Even with well-behaved integrands and sampling densities, the difficulty in approximating the baseline measure presents some unsolved problems which we discuss in Section 3.4.

### 3.2.1 Bridge sampling

Meng and Wong (1996) introduced the bridge sampling method for estimating ratios of normalizing constants and Kong et al. (2003) demonstrated its optimality using the likelihood perspective reviewed in the previous section. The main

example in this chapter extends the likelihood for the bridge sampling context identified by [Kong et al. \(2003\)](#), and we therefore now begin by describing bridge sampling and the setup used by [Kong et al. \(2003\)](#).

Suppose that we do not know  $c_s$ , for  $s = 1, 2$ , but that we have samples from  $p_1$  and  $p_2$  and want to estimate  $r = c_1/c_2$ . [Meng and Wong \(1996\)](#) propose computing an estimate  $\hat{r}$  using the following iterative scheme.

**Algorithm 1:** computing the bridge sampling estimate.

1. Choose  $\hat{r}^{(0)}$  and a positive integer  $T_{\max}$  (in practice,  $T_{\max}$  is usually sufficient).
2. Starting at  $t = 0$ , iteratively compute an estimate of  $r$  according to

$$\hat{r}^{(t+1)} = \frac{\frac{1}{n_2} \sum_{i=n_1+1}^n \frac{q_1(x_i)}{\frac{n_1}{n} q_1(x_i) + \frac{n_2}{n} \hat{r}^{(t)} q_2(x_i)}}{\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{q_2(x_i)}{\frac{n_1}{n} q_1(x_i) + \frac{n_2}{n} \hat{r}^{(t)} q_2(x_i)}}, \quad (3.6)$$

until convergence or  $t = T_{\max}$ . Here, the samples  $x_i$ , for  $i \in N_1 = \{1, \dots, n_1\}$ , are from  $p_1$  and the samples  $x_i$ , for  $i \in N_2 = \{n_1 + 1, \dots, n\}$ , are from  $p_2$ . For convenience we also write  $N = N_1 \cup N_2 = \{1, \dots, n\}$ .

The numerator and denominator of (3.6) are of the form  $\frac{1}{n_s} \sum_{N \setminus N_r} q_s(x_i) \alpha(x_i)$ , for  $s = 1$  and  $s = 2$ , respectively. Among estimators of this form, [Meng and Wong \(1996\)](#) and [Bennett \(1976\)](#) demonstrate that the bridge sampling estimator given by Algorithm 1 is asymptotically optimal in terms of relative mean squared error. Thus, it turns out that it is better to set the inverse weight  $(\alpha(x_i))^{-1}$  to be a mixture of the two sampling densities evaluated at  $x_i$ , rather than simply setting it

to be the density from which the sample  $x_i$  was drawn evaluated at  $x_i$ , which would be the choice if we estimated  $r$  by taking the ratio of two importance sampling estimates. The mixture form of the inverse weights explains the word “bridge” in bridge sampling: a stratified sample was indeed drawn from the mixture, and the mixture is being viewed as an importance sampling density that has more overlap with the two densities than they do with each other, i.e., it is a bridge between them.

[Kong et al. \(2003\)](#) assume the generalized scenario in which there are  $k \geq 2$  densities whose pairwise normalizing constant ratios are of interest and there are samples from at least one of the densities, with  $n_s$  denoting the number of samples from  $p_s$ , for  $s = 1, \dots, k$ . In this case, the likelihood is

$$L(\mu) = \prod_{i=1}^n p_{y_i}(x_i)\mu(x_i), \quad (3.7)$$

where  $y_i = s$  if the  $i^{\text{th}}$  sample is drawn from  $p_s$ , for  $s = 1, \dots, k$ . The parameter  $\mu$  is subject to the crucial constraints

$$\int_{\Omega} \frac{1}{c_s} q_s(x) d\mu(x) = 1, \quad (3.8)$$

for  $s = 1, \dots, k$ . Ignoring constants it follows that the log-likelihood is

$$l(\mu) = \sum_{i=1}^n \theta_i - \sum_{s=1}^k n_s \log c_s = n \int_{\Omega} \theta d\hat{P} - \sum_{s=1}^k n_s \log c_s, \quad (3.9)$$

where  $\theta_i = \log(\mu(x_i))$ , for  $i = 1, \dots, n$ , and  $\hat{P}$  denotes the empirical distribution

of the samples. Since the integral on the right hand side is actually a finite sum, we see that (3.9) has the form of an exponential family log density with canonical parameter  $\theta = (\theta_1, \dots, \theta_n)$  and canonical sufficient statistic  $\hat{P}$ . Exponential family theory tells us that we can estimate the canonical parameter by first setting the canonical sufficient statistic equal to its expectation, which in this case gives

$$\hat{P}(dx) = \sum_{s=1}^k \frac{n_s}{n} \frac{1}{\hat{c}_s} q_s(x) \hat{\mu}(dx), \quad (3.10)$$

and consequently the MLE of  $\mu$  is

$$\hat{\mu}(dx) = \frac{\hat{P}(dx)}{\sum_{s=1}^k \frac{n_s}{n} \frac{1}{\hat{c}_s} q_s(x)}. \quad (3.11)$$

Throughout we will simply write  $\hat{\mu}$  to denote the MLE of  $\mu$ . The corresponding MLEs of the normalizing constants are

$$\hat{c}_s = \int_{\Omega} q_s(x) d\hat{\mu}(x) = \sum_{i=1}^n \frac{q_s(x_i)}{\sum_{t=1}^k n_t \frac{1}{\hat{c}_t} q_t(x_i)}, \quad (3.12)$$

for  $s = 1, \dots, k$ . Finally, for any  $s, t \in \{1, \dots, k\}$ , we estimate  $r_{st} = c_s/c_t$  by

$$\hat{r}_{st} = \frac{\hat{c}_s}{\hat{c}_t}. \quad (3.13)$$

Simple algebra utilizing the constraints (3.8) verifies that this is the same as the bridge sampling estimator given by Algorithm 1 in the case where  $s = 1$ ,  $t = 2$ ,  $k = 2$ , and  $n_1, n_2 > 0$ . In all cases, the normalizing constants and  $\mu$  are only



estimated up to a common constant of proportionality, because the estimate of the measure depends on the normalizing constants and vice versa. Thus, only estimates of ratios such as  $r_{st}$ , for  $s, t \in \{1, \dots, k\}$ , are meaningful estimates.

As might be anticipated from the bridge sampling estimator computed by Algorithm 1,  $\hat{\mu}$  ignores the information about which density each sample  $x_i$  comes from. It is therefore apparent that it is the *constraints* given in (3.8) that allow us to “estimate” the measure, and the samples themselves only determine where  $\hat{\mu}$  should be non-zero. The fact that the MLE of the measure should only be non-zero at the sampled points was established by Vardi (1985), and in particular is because setting  $\hat{\mu}(x)$  to be zero at any point  $x$  that was not sampled minimizes the estimated normalizing constants  $\hat{c}_s$ , for  $s = 1, \dots, k$ , and hence maximizes the likelihood. We also suggest the work of Vardi (1985) to those who seek a more mathematically transparent approach to arriving at the MLE (3.11) since he derives a technically similar result to Kong et al. (2003) without invoking exponential family theory.

### 3.2.2 Asymptotic variance

Kong et al. (2003) and Kong et al. (2006) show that, under the maximum likelihood framework described in the previous section, the asymptotic covariance matrix for  $\log(\hat{c})$  can be calculated, where  $\hat{c} = (\hat{c}_1, \dots, \hat{c}_k)$ . Since the normalizing constants can only be estimated up to a common constant of proportionality, only contrasts of  $\log(\hat{c})$  have meaningful variance. Statistically, it is key to be able to approximate the variance of the estimates  $\hat{r}_{st}$ , for  $s, t \in \{1, \dots, k\}$ , and we there-

fore briefly outline a practical way to calculate the variance, but refer the reader to [Kong et al. \(2006\)](#) for further technical details since variance estimates are not crucial for understanding our main contribution.

By comparing (3.9) with a multinomial log-likelihood, or appealing to the Fisher information measure (see [McCullagh 1999](#)), it can be established that the infinite dimensional Fisher information matrix for  $\theta = \log \mu$  is

$$n\mathcal{I}_\theta(A, B) = \sum_{s=1}^k n_s(P_s(A \cap B) - P_s(A)P_s(B)), \quad (3.14)$$

where  $A, B \subset \Omega$ , and  $P_s(dx) = \frac{1}{c_s}q_s(x)\mu(dx)$ , for  $s = 1, \dots, k$ . The difference between (3.14) and the multinomial Fisher information is that the “categories”  $A$  and  $B$  may overlap. At  $\theta = \hat{\theta}$ , the matrix in (3.14) can be expressed in finite dimensional form because the MLE  $\hat{\theta}$  of the log measure has finite support. In particular, if  $\hat{P}$  is the  $n \times k$  matrix with  $(i, s)$  element

$$\hat{P}_s(x_i) = \frac{\frac{1}{c_s}q_s(x_i)}{\sum_{t=1}^k n_t \frac{1}{c_t}q_t(x_i)}, \quad (3.15)$$

and  $W = \text{diag}(n_1, \dots, n_k)$ , then the Fisher information matrix  $n\mathcal{I}_{\hat{\theta}}$  for  $\theta$  at  $\hat{\theta}$  is  $I_n - \hat{P}W\hat{P}^T$ , where  $I_n$  is the  $n \times n$  identity matrix. This can be verified by computing (3.14) for  $A = \{x_i\}$  and  $B = \{x_j\}$  for each pair  $(i, j) \in N \times N$ , and noting that in the case  $A = B = \{x_i\}$  we have  $\sum_{s=1}^k n_s(P_s(A \cap B)) = \sum_{s=1}^k n_s\hat{P}_s(x_i) = 1$ , for  $i = 1, \dots, n$ . It follows, albeit with some technical clarifications, that the

asymptotic covariance matrix of  $\log(\hat{c})$  is

$$\hat{V} = \hat{P}^T(I_n - \hat{P}W\hat{P}^T)^- \hat{P}. \quad (3.16)$$

The superscript “ $-$ ” denotes a generalized inverse matrix which in the current context is the inverse of  $I_n - \hat{P}W\hat{P}^T + \mathbf{1}_n\mathbf{1}_n^T/n$ , where  $\mathbf{1}_n$  is a vector of  $n$  ones (a generalized inverse of a matrix  $M$  is a matrix  $M^-$  such that  $MM^-M = M$ ). If  $k = 2$ , the asymptotic variance of  $\log(\hat{r}_{ML})$  is given by

$$\text{Var}(\log(\hat{r}_{ML})) = \hat{V}_{11} + \hat{V}_{22} - 2\hat{V}_{12}. \quad (3.17)$$

and we can of course convert to the variance of  $\hat{r}_{ML}$  by multiplying by  $\hat{r}_{ML}^2$ ,

### 3.2.3 Group invariance sub-models

It is possible to improve our estimates of  $r_{st}$ , for  $s, t \in \{1, \dots, k\}$ , by incorporating some of our knowledge of the true underlying measure  $\mu$ . In practice, we know the measure to be the Lebesgue measure, but if we invoke this knowledge completely then the MLE of the measure will no longer have finite support and we will again be faced with an intractable integral when we try to estimate normalizing constants (and their ratios). Instead, we must choose *some* knowledge of the true measure to incorporate in the likelihood, and in particular knowledge that improves the MLE approximation to the measure in regions where our unnormalized densities have large values.

One useful way of incorporating some knowledge about the measure is to con-

strain our estimate of  $\mu$  to be invariant to a group of transformations that we know  $\mu$  to be invariant to, e.g., the group of reflections about the origin. That is, if we can identify a compact group of transformations  $\mathcal{G}$  that we know satisfies the invariance  $\mu(A) = \mu(g(A))$ , for all  $g \in \mathcal{G}$  and  $A \subset \Omega$ , then we can incorporate this information when we write down the likelihood, and consequently  $\hat{\mu}$  will have the same invariance. The benefit of imposing this group invariance is that the new MLEs for the normalizing constants utilize all the points in the orbit  $\mathcal{G}(x_i) = \{g(x_i), g \in \mathcal{G}\}$  of each sample  $x_i$ , for  $i = 1, \dots, n$ , and therefore our “sample size” is larger by a factor of  $|\mathcal{G}|$ . Following [Kong et al. \(2003\)](#), we describe this approach as imposing a *group invariant sub-model* because, out of the collection of all possible measures, we are confining our attention to the smaller collection that have the specified invariance. Note that invariance simply means the *Jacobians* of the transformations  $g \in \mathcal{G}$  are equal to one, because a Jacobian specifies the way that the measure is altered by a transformation. Transformations with Jacobians equal to one have the important property that they leave normalizing constants unchanged.

[Kong et al. \(2003\)](#) incorporate group invariance of the measure into the likelihood by imagining that before seeing  $x_i$  we randomly select a transformation  $g \in \mathcal{G}$  (that is not observed) and apply it to  $x_i$  (with each  $g \in \mathcal{G}$  having equal probability of being selected). Thus, the observed data is  $g(x_1), \dots, g(x_n)$  and each sample  $g(x_i)$  is equally likely to have been generated from any of the  $|\mathcal{G}|$

densities of the form  $p_{y_i}(g(\cdot))$ , where  $g \in \mathcal{G}$ . Hence, the likelihood becomes

$$L(\mu) = \prod_{i=1}^n \frac{1}{c_{y_i}} \bar{q}_{y_i}(x_i) \mu(x_i), \quad (3.18)$$

under the constraints

$$\int_{\Omega} \frac{1}{c_s} \bar{q}_s(x) d\mu(x) = 1, \quad (3.19)$$

for  $s = 1, \dots, k$ , where

$$\bar{q}_s(x) \equiv \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} q_s(g(x)). \quad (3.20)$$

Importantly, the likelihood (3.18) utilizes the invariance  $\mu(x) = \mu(g(x))$ , for all  $g \in \mathcal{G}$ . Now, setting the canonical sufficient statistic equal to its expectation, the MLE of  $\mu$  is

$$\hat{\mu}(x) = \frac{1}{\sum_{s=1}^k n_s \frac{1}{c_s} \bar{q}_s(x)}, \quad (3.21)$$

if  $x \in \bigcup_{i=1}^n \mathcal{G}(x_i)$ , and zero otherwise. Equations (3.11) and (3.21) are the same except that (3.21) replaces  $q_s$  with  $\bar{q}_s$  and has larger support due to the imposed invariance, i.e.,  $\hat{\mu}(x) = \hat{\mu}(g(x))$  for all  $g \in \mathcal{G}$ . Lastly, since

$$c_s = \int_{\Omega} \bar{q}_s(x) d\mu(x), \quad (3.22)$$

the normalizing constant  $c_s$  is estimated by

$$\hat{c}_s = \sum_{i=1}^n \frac{\bar{q}_s(x_i)}{\sum_{t=1}^k n_t \frac{1}{\hat{c}_t} \bar{q}_t(x_i)}, \quad (3.23)$$

which is the same as (3.12) except that again  $\bar{q}_s$  has replaced  $q_s$ . An estimate of the variance of the corresponding  $\hat{r}_{st}$ , for  $s, t \in \{1, \dots, k\}$ , can similarly be obtained by replacing  $q_s$  by  $\bar{q}_s$  in (3.15).

The above approach of [Kong et al. \(2003\)](#) is very intriguing since it incorporates our knowledge of the invariance of the baseline measure into the likelihood by *discarding* the corresponding information in the data, i.e., by discarding the information about which of the densities  $\{p_{y_i}(g(\cdot)) : g \in \mathcal{G}\}$  generated sample  $i$ , for  $i = 1, \dots, n$ . This general principle of simultaneously inserting knowledge into the measure and taking information from the data could be very useful in further developments. Taking information from the data expands the sufficient statistic: in Section 3.2.1, the sufficient statistic was the empirical measure or simply the collection of samples, but now the sufficient statistic additionally includes the orbits of the samples. Intuitively, this expansion of the sufficient statistic corresponds to the fact that the empirical measure now has greater entropy or uncertainty. The corresponding *increased* knowledge or certainty about the measure, namely the group invariance constraint, means that the estimates (3.23) are essentially “Rao-Blackwellized” versions our previous estimates (3.12), see [Liu \(2008\)](#) Section 2.5.5. Importantly, this Rao-Blackwellization ensures that we cannot do worse by using invariances of the measure. (The alternative argument that the maximum

likelihood method ensures we cannot do worse is not quite valid in this case because we change the data, and in fact reduce the information they provide, when we impose invariances.) The guarantee of improvement is in contrast to alternative methods based on symmetries of the densities, rather than symmetries of the measure, because these approaches sometimes result in worse estimators, e.g., methods based on the antithetic principle.

We emphasize that, in the current context, improvements in estimation precision are due to additional density evaluations, and therefore come at a computational cost. Consequently, we are generally interested in small invariance groups, with the main examples being reflection and rotation groups, and permutation groups for discrete spaces. Small well chosen groups can potentially greatly increase precision at little additional computational cost (we are assuming evaluation of the densities is fast compared with generating additional samples). However, we note that some valuable transformation groups are excluded from the group invariant sub-model framework because they are infinite, e.g., translation groups (which are otherwise ideal since their transformation Jacobians are equal to one). Indeed, the reason we concentrate on the special case of groups is because otherwise repeatedly applying a transformation will generate an infinite number of points of invariance from just one sample. It can be argued that we only want to apply a given transformation once, but we must then find a way to impose this as a meaningful a priori constraint on the measure, which is the topic of Section 3.3.

### 3.3 Warp transformations and beyond

Location shifts and scale changes of the samples can greatly improve the performance of bridge sampling estimators, as demonstrated by [Meng and Schilling \(2002\)](#). The idea behind these transformations is to gain greater estimation precision by selectively *modifying* the sampled points to be used in integral approximations, rather than by evaluating integrands at more points, as is the idea under the group invariant sub-model framework discussed in the previous section. This sample modification method can intuitively be understood as “warping” the integrands and sampling densities into similar shapes, hence the name “warp transformations” introduced by [Meng and Schilling \(2002\)](#). In this section, we show that a simple augmentation of the samples places warp transformations under the group invariant sub-model framework, despite the apparent differences between the two approaches. This gives warp bridge sampling estimators a maximum likelihood interpretation, and allows us to improve estimation precision by both modifying the sample points and introducing orbits of the modified points, all under a single framework. Some essential details of warp bridge sampling will be mentioned but for a full exposition the reader is referred to [Meng and Schilling \(2002\)](#). For conciseness of the exposition, in the remainder of this chapter we will focus on the  $k = 2$  case and denote the estimand of interest and its MLE by  $r$  and  $\hat{r}_{ML}$ , respectively.



### 3.3.1 Warp I

The Warp I bridge sampling estimator translates the unnormalized densities before applying Algorithm 1, and in particular changes the iterative update (3.6) to

$$\hat{r}^{(t+1)} = \frac{\frac{1}{n_2} \sum_{i=n_1+1}^n q_1(x_i + D) \alpha(x_i)}{\frac{1}{n_1} \sum_{i=1}^{n_1} q_2(x_i - D) \alpha(x_i - D)}, \quad (3.24)$$

where

$$\alpha(x) = \frac{1}{n_1 q_1(x + D) + n_2 \hat{r}^{(t)} q_2(x)}. \quad (3.25)$$

Intuitively, if  $\mu_1$  and  $\mu_2$  are measures of the “center” of densities  $p_1$  and  $p_2$ , respectively, then we should choose  $D = \mu_1 - \mu_2$  (the subscripts distinguish  $\mu_1$  and  $\mu_2$  from the measure). The Warp I estimator does not immediately fall under the group invariant sub-model maximum likelihood framework of the previous section because, for one thing, the orbit generated by the translation  $g(x) = x + D$  is infinite. However, the difficulty can be overcome by using the label information that specifies which density each sample originated from. The label information naturally must play a key role in the warping context because the transformations are density specific. We augment the data space with the label information so that observations are  $\{x_i, y_i\}$ , for  $i = 1, \dots, n$ . Next, we introduce a new transformation group  $\mathcal{G}_D = \{I, g\}$  on the augmented space, where  $I$  is the identity

transformation, and

$$\begin{aligned} g(x, 1) &= \{x - D, 2\} \\ g(x, 2) &= \{x + D, 1\}. \end{aligned} \tag{3.26}$$

Here and throughout,  $g(x, y)$  is understood to mean the transformation  $g$  applied to the ordered pair  $\{x, y\}$ , i.e.,  $g(\{x, y\})$ . By using the label information to only allow alternating translations, we have reduced the infinite translation group to a cyclic group of order two! Furthermore, this choice exactly recovers the Warp I transformation under the sub-model theory as we now show. The definition of the density  $p_s$  on the augmented space is

$$p_s(x, y) = \frac{1}{c_s} q_s(x, y) \equiv \begin{cases} \frac{1}{c_s} q_s(x) & y = s \\ 0 & y \neq s, \end{cases} \tag{3.27}$$

for  $s = 1, 2$ , because we know that  $\{x, y\}$  cannot be from  $p_s$  unless  $s = y$ . Here and throughout, it is convenient to let  $q_s$  denote both the function acting on  $\{x, y\}$  and the function only acting on  $x$ , and there is no confusion because we always specify the argument when distinction is necessary. The group invariant sub-model framework now tells us that MLEs of the normalizing constant are given by (3.23), and since  $k = 2$ , we have

$$\hat{r}_{ML} = \frac{\hat{c}_1}{\hat{c}_2} = \frac{\sum_{i=1}^n \frac{\bar{q}_1(x_i, y_i)}{n_1 \bar{q}_1(x_i, y_i) + n_2 \hat{r}_{ML} \bar{q}_2(x_i, y_i)}}{\sum_{i=1}^n \frac{\bar{q}_2(x_i, y_i)}{n_1 \bar{q}_1(x_i, y_i) + n_2 \hat{r}_{ML} \bar{q}_2(x_i, y_i)}}. \tag{3.28}$$

which is exactly the Warp I bridge sampling estimate (3.24) upon convergence (after some simple algebra verifying that summing over all  $n$  in (3.24) would not change the value of  $\hat{r}^{(t+1)}$ ). Note that, since we have modified the samples rather than generating additional evaluation points, Warp I bridge sampling does not have Rao-Blackwellization interpretation discussed in Section 3.2.3 and its optimality is conditional on the translation parameter  $D$ . We discuss methods for choosing  $D$  in Section 3.4. In the next section, we apply our formulation to generalized warp transformations, which have the additional complication that their Jacobians may not be equal to one.

### 3.3.2 Generalized warping

Our augmentation approach used to place translations under the group invariant sub-model framework is very convenient because it works for any invertible transformation of the samples. We now take the remaining step needed to verify this, namely to incorporate transformations whose Jacobians are not equal to one. We illustrate with Warp II transformations, which center and scale densities, and Warp III transformations which additionally symmetrize densities.

Consider the group  $\mathcal{G}_T = \{I, g\}$  with non-identity element

$$\begin{aligned} g(x, 1) &= \{T(x), 2\} \\ g(x, 2) &= \{T^{-1}(x), 1\}. \end{aligned} \tag{3.29}$$

To accommodate instances of  $T$  whose Jacobians are not equal to one, we gener-

alize the definition of group averaging given in (3.20), at least for the case where  $k = 2$  and the invariance group has order two. Specifically, we set

$$\bar{q}_s(x, y) \equiv \frac{1}{2}q_s(x, y) + \frac{1}{2}q_s(g(x, y))|J_s(x)|, \quad (3.30)$$

for  $s = 1, 2$ , where  $J_1$  is the Jacobian of  $T^{-1}$  and  $J_2(x) = (J_1(x))^{-1}$ . For any  $x$ , one of the two terms in (3.30) is zero due to the definition of  $q_s$ . To obtain  $\hat{r}_{ML}$ , we now simply use (3.30) in place of  $q_s(x)$  in Algorithm 1, for  $s = 1, 2$ . For clarity, we denote the resulting estimator by  $\hat{r}_{ML}(\mathcal{G}_T)$  and will use similar notation for other estimators.

In the case of deterministic transformations, [Meng and Schilling \(2002\)](#) give a general form of their warp bridge sampling method in which the two densities are both warped to a standard shape (e.g., a Normal distribution) by the transformations  $H_1$  and  $H_2$ , yielding the final estimator  $\hat{r}_{GW}(H_1, H_2)$  (where GW stands for generalized warp bridge sampling estimator). This generalized warp bridge sampling estimator is the solution to

$$\hat{r} = \frac{\frac{1}{n_2} \sum_{i=n_1+1}^n q_1(H_1^{-1}(H_2(x_i)))|J_{H_1^{-1}}(H_2(x_i))|\alpha(H_2(x_i))}{\frac{1}{n_1} \sum_{i=1}^{n_1} q_2(H_2^{-1}(H_1(x_i)))|J_{H_2^{-1}}(H_1(x_i))|\alpha(H_1(x_i))}, \quad (3.31)$$

where  $J_{H_s^{-1}}$  is the Jacobian of  $H_s^{-1}$  and

$$\alpha(x) = \frac{1}{n_1 q_1(H_1^{-1}(x))|J_{H_1^{-1}}(x)| + n_2 \hat{r} q_2(H_2^{-1}(x))|J_{H_2^{-1}}(x)|}. \quad (3.32)$$

If we choose  $T = H_2^{-1}H_1$  in (3.29), then carefully accounting for Jacobian terms

gives  $\hat{r}_{GW}(H_1, H_2) = \hat{r}_{ML}(\mathcal{G}_T)$ , see Appendix A.5. Thus, in the case of deterministic transformations, all warp bridge sampling estimators can be expressed as maximum likelihood estimators under an appropriate group invariant sub-model.

The Warp II transformation centers and scales the two densities and therefore is a specific example in which the corresponding transformation  $T$  in (3.29) will not have a Jacobian equal to one. In [Meng and Schilling \(2002\)](#), the density specific Warp II transformations are given by

$$H_s(x) = S_s^{-1}(x - \mu_s) \quad (3.33)$$

where  $\mu_s$  and  $S_s$  are measures of the center and spread of the density  $p_s$ , for  $s = 1, 2$ . Therefore, under our sample augmented group invariant sub-model approach, the corresponding transformation  $g$  is

$$\begin{aligned} g(x, 1) &= \{Sx + D, 2\} \\ g(x, 2) &= \{(x - D)/S, 1\}, \end{aligned} \quad (3.34)$$

where  $S = S_2 S_1^{-1}$  and  $D = \mu_2 - S\mu_1$ . The jacobians used in the group averaging (3.30) are  $J_1(x) = S$  and  $J_2(x) = S^{-1}$ , for all  $x \in \Omega$ . We denote  $\{I, g\}$  and the resulting estimator by  $\mathcal{G}_{W2}$  and  $\hat{r}_{ML}(\mathcal{G}_{W2})$ , respectively.

Warp III transformations make densities symmetric by using

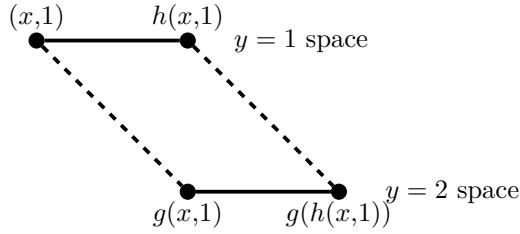
$$\bar{q}_s(x) = \frac{1}{2}(q_s(x) + q_s(-x)) \quad (3.35)$$

**Table 3.1:** Cayley table of the invariance group  $\mathcal{G}_{W3}$  used by the sub-model of the measure that corresponds to Warp III transformations.

	$I$	$g$	$h$	$gh$
$I$	$I$	$g$	$h$	$gh$
$g$	$g$	$I$	$gh$	$h$
$h$	$h$	$gh$	$I$	$g$
$gh$	$gh$	$h$	$g$	$I$

in place of  $q_s(x)$  in Algorithm 1, for  $s = 1, 2$ . Meng and Schilling (2002) point out that this can be interpreted as bridge sampling estimation following a stochastic transformation of the samples, namely multiplying each by an independent random sign. From the maximum likelihood perspective, the use of (3.35) is clearly a direct application of the group averaging method reviewed in Section 3.2.3; there is no need for our sample augmentation. However, to obtain the best results we would usually combine (3.35) with the Warp II transformation, which does require augmented samples for the group invariant sub-model interpretation. Specifically, the corresponding invariance group is  $\mathcal{G}_{W3} = \{I, g, h, gh\}$ , where  $g$  is given by (3.34) and  $h$  is the same except that the parameter  $D$  is replaced by  $D' = \mu_2 + S\mu_1$ . Note that,  $gh(x, y) = hg(x, y) = \{-x + 2\mu_y, y\}$  and therefore  $gh$  is simply a reflection about the center  $\mu_y$  of the density  $p_y$ , for  $y = 1, 2$ . Thus, it is easily verified that  $\hat{r}_{W3}(H_1, H_2) = \hat{r}_{ML}(\mathcal{G}_{W3})$  through an alternative calculation of  $\hat{r}_{ML}(\mathcal{G}_{W3})$ : first replace  $q_s(x, y)$  by

$$\check{q}_s(x, y) = \frac{1}{2} (q_s(x, y) + q_s(gh(x, y))) = \frac{1}{2} (q_s(x, y) + q_s(-x + 2\mu_s, y)), \quad (3.36)$$



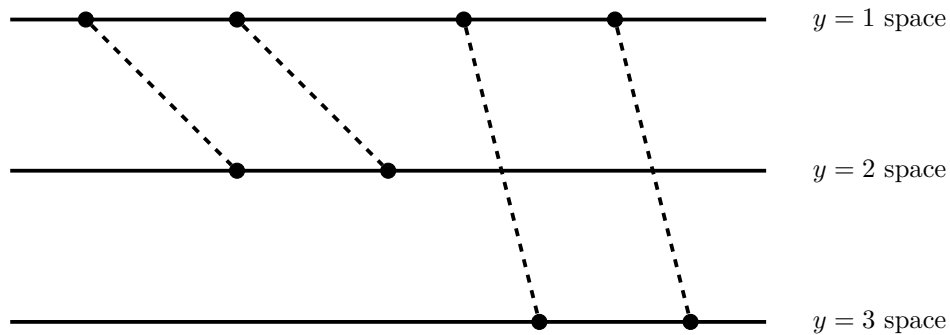
**Figure 3.1:** Orbit of  $(x,1)$  under the invariance group  $\mathcal{G}_{W3}$  that corresponds to Warp III transformations.

for  $s = 1, 2$ , and then compute  $\hat{r}_{ML}(\mathcal{G}_{W2})$  (with  $g$  given in (3.34)).

The group  $\mathcal{G}_{W3}$  is illustrated in Table 3.1 and Figure 3.1, and is helpful for gaining insight regarding further developments. It is straightforward to introduce more densities (that we may or may not sample from) by allocating some to live on the  $y = 1$  space and some to live on the  $y = 2$  space for the purposes of transformation and group averaging. However, in pursuing other developments, we must take note of the fact that the dashed lines in Figure 3.1 are parallel, since without this feature the transformations  $g$  and  $h$  could be combined to form an infinite orbit. For example, this means that expanding the invariance group to include further transformations must be done with care and the necessary modifications are discussed in the next section.

### 3.3.3 Larger invariance groups

When there are only two densities of interest the invariance groups can easily be extended by adding more transformations that map within the  $y = 1$  or  $y = 2$  space and do not create infinite orbits, e.g., orthogonal reflections. The within



**Figure 3.2:** Illustration of a sample augmentation that facilitate multiple cross space transformations in a group invariant sub-model.

space transformations do not have to be the same for the two spaces, and indeed they are not for the  $\mathcal{G}_{W_3}$  group. Through the group averaging, these within space transformations increase the precision of the maximum likelihood estimator by increasing the number of density evaluation points.

Cross space transformations are more complex and there can only be one such transformation under the two space augmentation. A second transformation would not be “parallel” (see Figure 3.1), and therefore the orbit of each sample would become infinite because we could traverse a space indefinitely by alternatingly applying the cross space transformations. Furthermore, if there are only two densities there would be no need for another cross space transformation; the one transformation we have should be our best attempt to transform one density to the other. It is possibly of interest to consider several cross space transformations in an attempt to include the optimal one, though this is likely to be inefficient. In the next section more will be said about how to choose the transformations.



If we have more than two densities, then further cross space transformations will be needed to make the most of the group averaging. To illustrate, suppose there are three unnormalized densities, with  $q_1$  living on  $y = 1$  and  $q_2$  and  $q_3$  living on  $y = 2$ . The points from space  $y = 1$ , say, will not map to points of space  $y = 2$  for which both  $q_2$  and  $q_3$  have high values unless they happen to overlap a lot. Considering this it becomes apparent that it is natural to introduce one space per density. Then, we can link each space to  $y = 1$  through a transformation, warping density  $p_s$  (on space  $y = s$ ) to density  $p_1$  as accurately as possible, for  $s = 2, \dots, k$ . This extension is illustrated in Figure 2. The result is that for any two densities  $q_s$  and  $q_t$ , we can approximately transform between them by first transforming  $q_s$  to  $q_1$ , and then transforming  $q_1$  to  $q_t$ . As we have already shown this is essentially equivalent to the idea in [Meng and Schilling \(2002\)](#) of transforming every density to look like a standard Normal density, except that here we use  $q_1$  as the reference rather than the standard Normal.

### 3.4 Choosing invariance group transformations

It desirable to optimize the invariance group transformations because a smaller group is both computationally cheaper and often more interpretable. We should be careful to make optimization fast so that it does not require as much computation time as just expanding the invariance group in a non-optimal way or drawing more samples. It is sufficient to optimize the transformations using a reasonably sized subset of the samples and then proceed to estimating  $r$  with all the samples. If

the sample size was large, this would clearly compare favorably with using a large non-optimized invariance group because for large sample sizes the cost of group averaging becomes important.

Section 3.4.1 provides a practical decision-theoretic approach to choosing invariance group transformations, that follows the method that [Meng and Schilling \(2002\)](#) suggested for choosing warping parameters. Despite the fact that this approach works well, it is somewhat unsatisfying because it breaks from the likelihood framework, and thus, may not correspond to optimal joint estimation of the transformation parameters and the measure. This may lead to sub-optimal estimation of  $r$  and leaves us uncertain if further development of estimators is warranted. It is therefore of interest to formulate a complete (possibly hierarchical) model under which the maximum likelihood approach can be used to jointly estimate the measure and the optimal transformation parameters. The “reverse mixture” likelihood based method explained in Section 3.4.3 offers a first step and appears to be only slightly inferior to the decision theoretic approach. Of course, a complete likelihood approach would not be inferior and may be superior. By presenting the reverse mixture based method, and other likelihood ideas, we aim to create a clearer understanding of the difficulties to encourage and facilitate further investigations.

### 3.4.1 Decision theoretic approach

Theorem 1 of [Meng and Wong \(1996\)](#) gives the asymptotic relative mean square error (RME) of the bridge sampling estimator as

$$\frac{E[(\hat{r} - r)^2]}{r^2} \stackrel{\cdot}{=} \frac{1}{n} \left[ \int_{\Omega_1 \cap \Omega_2} \left( \frac{1}{\frac{n_1}{n} p_1(x)} + \frac{1}{\frac{n_2}{n} p_2(x)} \right)^{-1} dx \right]^{-1} - \frac{1}{n_1} - \frac{1}{n_2}. \quad (3.37)$$

The dot over the equals sign indicates this is a first order equality and thus asymptotically an equality in the usual sense that the ratio of the two sides converges to 1. An essentially equivalent result was given in the physics literature by [Bennett \(1976\)](#).

Given the above results, one way to choose warp bridge sampling parameters is to minimize the (scaled) harmonic distance

$$\left[ \int_{\Omega_1 \cap \Omega_2} \frac{p_1(x)p_2(x)}{n_1 p_1(x) + n_2 p_2(x)} dx \right]^{-1} \quad (3.38)$$

once we have replaced the densities with their warped counterparts. Exactly the same approach can be applied for choosing any invariance group transformations; we can simply replace  $q_r$  with  $\bar{q}_r$  and minimize the harmonic distance. Of course, it is impossible to calculate the harmonic distance  $H$  without knowing the normalizing constants, but we can estimate the harmonic distance, up to a constant, by

$$\hat{H}_{\hat{r}}(\psi) = \sum_{i=1}^n \frac{\bar{q}_1(x_i, y_i) \bar{q}_2(x_i, y_i)}{(n_1 \bar{q}_1(x_i, y_i) + n_2 \hat{r} \bar{q}_2(x_i, y_i))^2}, \quad (3.39)$$

where  $\hat{r}$  is our current estimate of  $r$ , and  $\psi$  denotes the invariance group parameters. Thus, denoting the invariance group corresponding to  $\psi$  by  $\mathcal{G}_\psi$ , the following algorithm can be used to estimate  $r$ :

**Algorithm 2 (RME method):** decision-theoretic choice of invariance group parameters.

1. Choose an initial  $\hat{r}$  and  $\hat{\psi}$  and then iterate steps (1) and (2) until convergence.
2. Calculate  $\hat{r}$  using the scheme in (3.6) with  $q_s$  replaced by  $\bar{q}_s$ , for  $s = 1, 2$ , where the group averaging is over  $\mathcal{G}_{\hat{\psi}}$  (we must also replace  $x_i$  with  $\{x_i, y_i\}$ ).
3. Update  $\hat{\psi}$  by minimizing  $\hat{H}_{\hat{r}}(\psi)$  with respect to  $\psi$ .

To save computation we could begin by optimizing with respect to a small group, such as the Warp I translations, in order to obtain a reasonably good estimate of  $\hat{r}$ . Then, we could run a few iterations of Algorithm 2 beginning at step (3) and iterating between steps (2) and (3). This method is likely to work quite well in most cases because  $\hat{r}$  will not change massively once the densities are aligned. Indeed, the diminishing return of adding further transformations mean it is generally not a good idea to use groups that are very large.

Since the proposed  $\hat{\psi}$  does not clearly correspond to the MLE (see the the next section), we do not know if the above scheme yields an optimal  $\hat{r}$ . Simply arguing that we chose  $\hat{\psi}$  to minimize the asymptotic relative error of  $\hat{r}$  is not sufficient because this was only done approximately; the harmonic distance  $H$  is not known exactly. The general method implemented can be summarized as choosing  $\psi$ ,

the nuisance parameter, to maximize the Fisher information (or minimize the asymptotic variance) of  $\mu$ , the parameter of interest. In general, this does not provide the best estimate of  $\mu$  unless the parameters  $\psi$  and  $\mu$  are orthogonal. Furthermore it is not clear what is the best way to adapt Algorithm 2 to more than two sampling densities, and a complete maximum likelihood formulation may shed some light on this problem. It would also have the advantage of using the sufficient statistic which would potentially yield better estimates of  $r$  and  $\psi$  for finite sample sizes, though this is not guaranteed.

These issues aside, Algorithm 2 provides the best method we currently have for selecting invariance group parameters and it is easy to implement. In a test example, using a Normal and Skewed-Normal as the two sampling densities, Algorithm 2 provided noticeably better results than cruder approaches such as choosing a translation parameter by matching modes of the two densities. We mention in passing that [Meng and Schilling \(2002\)](#) also developed bounds for the asymptotic relative error of  $\hat{r}$  in terms of the Hellinger distance. Thus, we could update  $\hat{\psi}$  by minimizing the upper bound if this was for some reason computationally more practical than step (3) of Algorithm 2.

### 3.4.2 Difficulties with the ML approach

Intuitively, we would like to use the likelihood framework to jointly estimate the invariance group parameters and the measure. However, this idea has many difficulties rooted in the fact that the MLE of the measure turns out to be a function of the MLE of the invariance group parameters. For example, in the Warp I case,

we have  $\hat{\mu} = (\hat{D}, \hat{u})$ , where  $\hat{u}$  tells us the weight the estimated measure assigns to each of the observed points and  $\hat{D}$  tells us the (translational) invariance property of the estimated measure. We intuitively imagine  $\hat{D}$  to be telling us about the densities and, in particular, how best to transform between them, but it is actually telling us about the measure. This is on reflection unsurprising because the measure is the only unknown appearing in the likelihood (3.18).

For the remainder of this section, we will restrict attention to the Warp I case of translations (i.e.,  $g$  is given by (3.26)), but the discussion is relevant to the general formulation detailed in Section 3.3. For the Warp I case, we now mathematically demonstrate that attempting joint estimation by directly applying the likelihood approach discussed so far is unfruitful. From (3.18) and (3.26), the likelihood is

$$L(D, w) = \prod_{i=1}^{n_1} \frac{1}{c_1} q_1(x_i, y_i) u_i \prod_{i=n_1+1}^n \frac{1}{c_2} q_2(x_i, y_i) u_i, \quad (3.40)$$

where

$$c_1 = \sum_{i=1}^{n_1} q_1(x_i, 1) u_i + \sum_{i=n_1+1}^n q_1(x_i + D, 1) u_i, \quad (3.41)$$

and similarly for  $c_2$ . Now, suppose that the MLE is  $(\hat{D}, \hat{u})$ , and consider the corresponding normalizing constants  $\hat{c}_s$ , for  $s = 1, 2$ . Increasing the magnitude of  $\hat{D}$  sufficiently will decrease the estimates of  $\hat{c}_s$ , for  $s = 1, 2$ , and therefore increase the likelihood; a contradiction. Thus, the likelihood cannot be maximized with finite  $\hat{D}$ . If the unnormalized density  $q_2 g$ , say, has no mass at  $\{x_i, 1\}$ , for  $i = 1, \dots, n_1$ , this allows higher likelihood because then the *constraint* on the integral

of  $\bar{p}_2$  (which must be equal to one) does not restrict our choice of  $\mu(x_i, 1)$ , for  $i = 1, \dots, n_1$ . Thus, the likelihood chooses  $D$  such that  $q_s g$  has very little mass at  $\{x_i, s\}$ , and  $\hat{\mu}(x_i, s) = \hat{\mu}(g(x_i, s)) \approx \hat{c}_s / (n_s q_s(x_i, s))$ , for  $i \in N_s$  and  $s = 1, 2$ . The maximum likelihood approach therefore recovers the ratio of importance sampling estimators, which we know to be inferior to Warp I bridge sampling, at least for some choices of  $D$ .

The above arguments are difficult to reconcile because we expect the maximum likelihood approach to yield the optimal joint estimation procedure that we desire. It is tempting to think that the problem lies in the fact that the measure is an infinite dimensional parameter and therefore cannot be estimated. However, even for densities and measures with a finite support, we can still increase the likelihood by choosing  $D$  to minimize, rather than maximize,  $q_s g$  at the points  $\{x_i, s\}$ , for  $i \in N_s$  and  $s = 1, 2$ . For fixed  $D$ , the mixture form of the maximum likelihood estimator of the measure is not clear from the “likelihood” itself, but rather from the *constraints* on the measure, i.e., from the fact that the densities must integrate to one. Indeed, the bridge sampling estimator is of a different form to importance sampling estimator because there are more restrictions, which give us additional information about the measure. As we have seen, when we allow  $D$  to vary, this information is lost and  $\hat{\mu}$  is again the importance sampling estimator. In general, the warp bridge sampling estimator is not necessarily better than the ratio of importance sampling estimators, but for a good choice of  $D$  it is. In conclusion, the trouble is that what constitutes a good choice of  $D$  is not currently found anywhere in the data generating model, but is only realized through a study of  $\hat{r}$

itself.

Vardi (1985) investigates a closely related topic, namely the conditions required for the non-parametric MLE of a CDF  $F$  to exist in the scenario where samples are drawn from multiple weighted versions of  $F$ , i.e., when the sampling is biased. The  $k$  different weighted sampling schemes are described by the CDFs

$$F_s(t) = W_s(F)^{-1} \int_{-\infty}^t w_s(u) dF(u) \quad s = 1 \dots, k, \quad (3.42)$$

with

$$W_s(F) = \int_{-\infty}^{\infty} w_s(u) dF(u) \quad s = 1 \dots, k. \quad (3.43)$$

Vardi (1985) shows that for the MLE to exist, any subset  $B$  of the sampling densities must contain an  $s \in B$  such that  $w_s$  has non-zero weight for at least one of the samples drawn from the remaining densities not in  $B$ . This condition essentially means that there is sure to be some “empirical overlap” whichever way you divide up the sampling densities into groups  $B$  and  $B^C$ , i.e., at least some of the samples that came from one group of densities *could* have come from the other group. In our case, there is no direct overlap between the densities because of the space augmentation, and the overlap through the invariance group transformations is lost when we estimate the transformations, i.e., when the transformations are not fixed. We therefore need to change the *knowledge* we input into the likelihood in a similar fashion as we did when introducing group invariances of the measure in Section 3.2.3. Indeed, the input knowledge can dramatically influence the part



$D$  plays, as illustrated in the next section.

### 3.4.3 Reverse mixture method

We now investigate the method of *discarding* the information about whether a sample was generated from  $\bar{p}_1$  or  $\bar{p}_2$ , because this ensures that the resulting likelihood must contain information to estimate the “distance” between the densities, i.e.,  $D$ . Usually we want to disentangle components of a mixture distribution, but here we want to make them as similar as possible and we therefore call our approach the *reverse mixture method*. The method is slightly inferior to the decision theoretic approach of the previous section, but it illustrates the powerful role of the information we *select* to include in the likelihood and can potentially motivate further ideas. After discarding the sample label information, the Warp I likelihood is

$$L(\mu, D) = \prod_{i=1}^n \left( \frac{n_1}{n} \frac{1}{c_1} \bar{p}_1(x_i, y_i) + \frac{n_2}{n} \frac{1}{c_2} \bar{p}_2(x_i, y_i) \right) \mu(x_i, y_i). \quad (3.44)$$

This likelihood is appealing for the following reasons: (i) it appears that the maximum likelihood method must optimize  $g$  to map one density to the other; (ii) maximizing the likelihood has a nice interpretation as minimizing entropy of the empirical mixture; (iii) the estimate of the measure was not using the label information anyway so it seems we have recovered information about the “real”  $D$  at no cost. Unfortunately there are two objections that dramatically weaken the approach: (i) equation (3.44) assumes the real  $g$  transforms between the densities

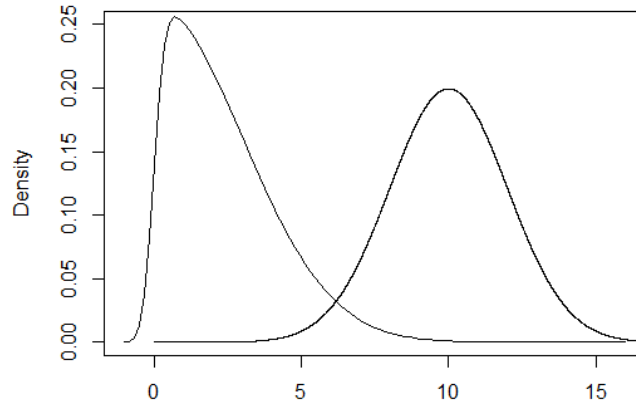
exactly; (ii)  $D$  is not identifiable.

Regarding problem (i), the issue is that  $\{\bar{x}_i, \bar{y}_i\}$  is not a sample from the density  $\frac{n_1}{n}\bar{p}_1 + \frac{n_2}{n}\bar{p}_2$  unless  $g$  perfectly transforms between  $p_1$  and  $p_2$ . In particular, if  $(x_i, 1)$  is drawn from  $p_1$  then it is not true that  $g(x_i, 1)$  is a draw from  $p_2$ , which is what we are assuming in (3.44). If we considered a sufficiently flexible transformation  $g$  (with more parameters) then the assumption would likely be a reasonable approximation. However, the parameters would still be non-identifiable because the MLE of  $\mu(x_i, y_i)$  is simply the reciprocal of  $n_1 \frac{1}{c_1} \bar{q}_1 + n_2 \frac{1}{c_2} \bar{q}_2$  and the maximum likelihood is equal to one for any transformation parameters, i.e., we are in the importance sampling situation of Section 3.2. This difficulty is also possible to overcome because we know the true measure. We propose the following scheme:

**Algorithm 3 (RM method):** reverse mixture method for choosing invariance group parameters.

1. First choose an initial  $\hat{r}$  and  $\hat{\psi}$  and then iterate steps (2) and (3).
2. Calculate  $\hat{r}$  using the scheme in (3.6) with  $q_s$  replaced by  $\bar{q}_s$ , for  $s = 1, 2$ , where the group averaging is over  $\mathcal{G}_{\hat{\psi}}$  (we must also replace  $x_i$  with  $(x_i, y_i)$ ).
3. Multiply (3.44) by  $c_1$  and insert  $\hat{r}$  in place of  $r$ . Then find the MLE of  $\psi$  while fixing  $\mu$  to be the true measure, i.e., minimize the entropy of the empirical mixture.

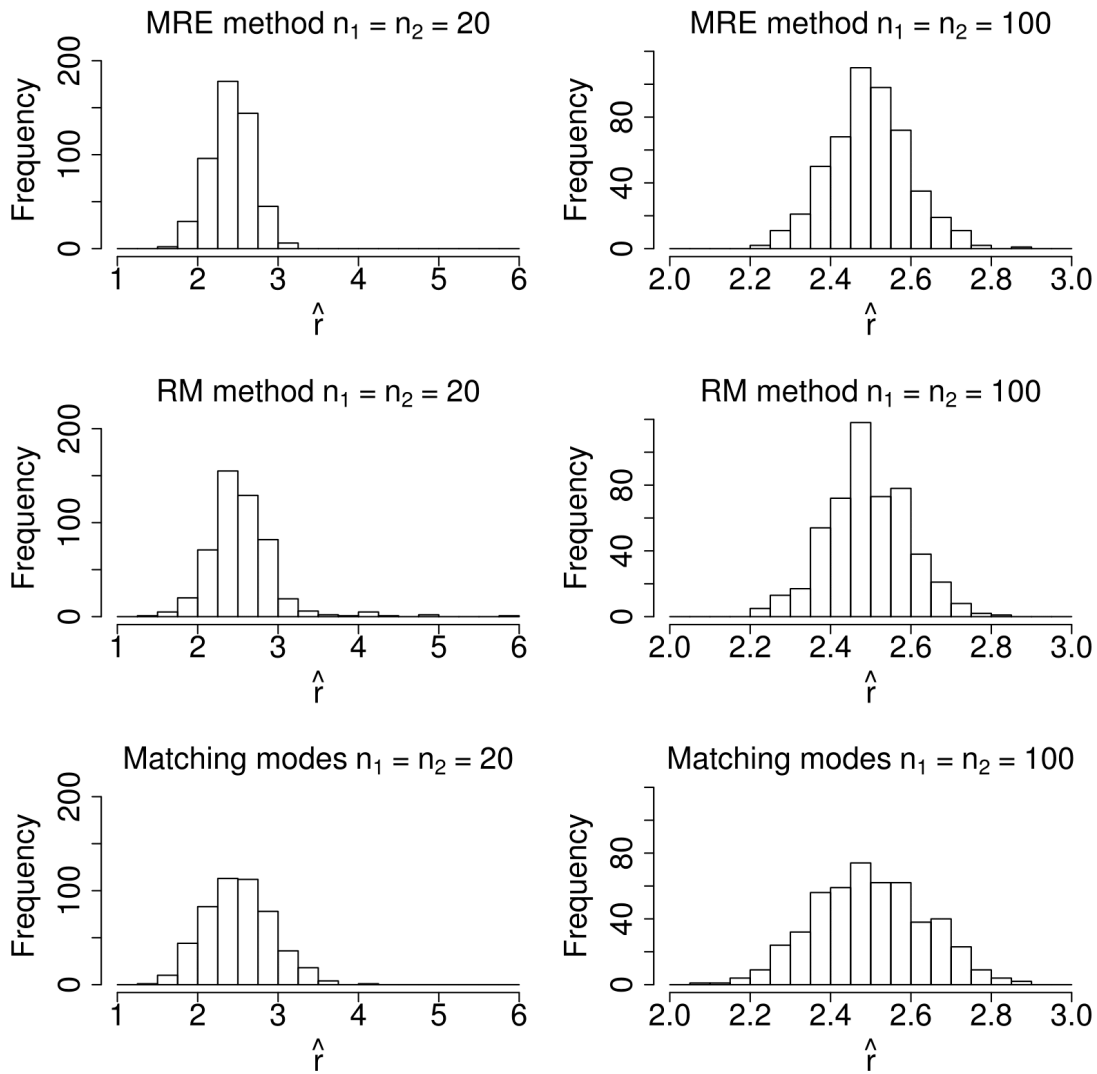
In the case  $\psi = D$ , this scheme produces similar estimates of  $\hat{r}$  as the MRE method of Section 3.4.1, but results in slightly higher mean squared error (see Figure 3.4, discussed below). Algorithm 3 is likely to perform better as we increase the



**Figure 3.3:** Plots of the  $N(10, 2)$  and Skew-Normal(0, 3, 10) probability densities used in our simulation study.

number of transformation group parameters, because then the assumption that  $g$  transforms between the densities exactly will become more reasonable. Therefore, when using Warp I, II, and III instead of only Warp I we may expect the MRE and RM methods to closely agree. Furthermore, the RM method does have one practical advantage over the decision-theoretic approach in that it has an intuitive generalization to more sampling densities; in step (3) of Algorithm 3 we simply minimize the entropy of the empirical mixture of all the sampling densities (step (1) and (2) also require minor alterations). Finally, it should be noted that, asymptotically the two approaches give very slightly differing values of  $D$  because we are optimizing entropy in one case and harmonic distance in the other.

We now give a representative example of how the suggested methods compare when choosing the translation parameter  $D$ . We set  $p_1$  to be a  $N(10, 2)$  density and  $p_2$  to be a Skew-Normal(0, 3, 10) density, see Figure 3.3 for plots of these densities. The  $r$  was set to 2.5, and we estimated it using the maximum likelihood



**Figure 3.4:** Estimates of  $r$  across the 500 simulations; the rows correspond to the three methods of choosing  $D$  (described in the main text) and the columns correspond to the two simulation settings  $n_1 = n_2 = 20$  (left) and  $n_1 = n_2 = 100$  (right).

Warp I estimator given in (3.28) with  $D$  chosen using three different methods: (i) the decision-theoretic method (MRE), (ii) the reverse mixture method (RM), and (iii) matching the true modes of the densities. The simulation was performed 500 times for each of the two settings  $n_1 = n_2 = 20$  and  $n_1 = n_2 = 100$ . Each panel in Figure 3.4 shows a histogram of the estimates of  $r$  across the 500 simulations; the rows correspond to the three methods of choosing  $D$  and columns correspond to the two simulation settings. As can be seen from the spread of histograms, the MRE approach did best in both simulations, though is only marginally better than the RM method for the larger sample sizes. The mean squared errors corresponding to the MRE, RM, and matching modes methods under  $n_1 = n_2 = 20$  were 0.078, 0.182, and 0.171, respectively, and under  $n_1 = n_2 = 100$  they were 0.0098, 0.0101, and 0.0192, respectively. The matching modes approach cheats slightly in that it uses the true modes of the densities which would be unknown in most applications. It is therefore perhaps unsurprising that the mode matching approach works slightly better than the RM method for the small sample simulation, but compares poorly when the sample size is increased.

### 3.5 Summary and future work

In this chapter we demonstrate that the Warp bridge sampling methods proposed by [Meng and Schilling \(2002\)](#) fall under the likelihood framework for Monte Carlo integration introduced by [Kong et al. \(2003\)](#) and [Kong et al. \(2006\)](#). Previously, [Kong et al. \(2003\)](#) and [Kong et al. \(2006\)](#) showed that statistical efficiency could

be gained by selectively inserting our knowledge about invariances of the baseline measure into the likelihood of the Monte Carlo samples, and in particular by use of the group invariant sub-models for the measure reviewed in Section 3.2.3. Statistical efficiency is gained through evaluating the integrand at more points introduced by the invariances of the measure, but importantly the method does not require additional samples, and thus is often possible even when only minor increases in computation time are acceptable. Rotations and reflections were the main invariances that could previously be exploited, but we introduce a simple augmentation of the Monte Carlo samples and measure that allows any one-to-one transformation for which we can compute the Jacobian to be used in the group invariant sub-model framework. This greatly increases the power of the sub-model framework and shows that Warp bridge sampling methods have a maximum likelihood interpretation. We also suggest both a decision-theoretic and a likelihood based method for choosing the optimal invariances within a given class.

Our future work will focus on identifying a likelihood that allows the normalizing constants (or other integrals) and the optimal sub-model parameters to be jointly estimated, since neither of our current approaches yet does this satisfactorily. The successful method will achieve *approximate* invariance of the measure based on the *observed* samples, as opposed to the exact invariance imposed in the likelihood (3.40) between the observed samples and the hypothetical samples in their orbits. Since approximate invariance is required, a hierarchical model could be useful for including our knowledge of the measure, although the specifics are not yet clear to us. Potentially related is the seminal work of [Efron and Morris \(1975\)](#)

who point out that the James-Stein estimator, which initially seems to be at odds with usual maximum likelihood theory, is natural given an appropriate hierarchical model. Indeed, some analogies can be made between our samples and the half season and full season batting averages in the infamous baseball example used by [Efron and Morris \(1975\)](#).

Another intriguing direction for future work is to identify conditions under which it is possible to estimate a function of an infinite dimensional parameter consistently. There will need to be conditions on both the sampling method and the type of function that is being estimated. In our case the function is integration with respect to the infinite dimensional measure. We need to invoke some form of smoothness and also some form of concentration of “mass” otherwise a finite approximation to the measure will not allow accurate estimation. An appropriate theorem would provide a stronger foundation for the likelihood approach to Monte Carlo integration, because for example so far it has not been rigorously demonstrated that the likelihood approach of [Kong et al. \(2003\)](#) and [Kong et al. \(2006\)](#) leads to consistent estimators.

# A

## Appendices

### A.1 Proof of Theorem 1.1

A Taylor expansion in  $\theta_0$  shows that  $\mathcal{V}(f(x|\theta_0)/f(x|\theta_1)) - \mathcal{V}(1)$  is equal to

$$\left( \delta \frac{f'(x|\theta_1)}{f(x|\theta_1)} + \frac{\delta^2}{2} \frac{f''(x|\theta_1)}{f(x|\theta_1)} \right) \mathcal{V}'(1) + \frac{\delta^2}{2} \left( \frac{f'(x|\theta_1)}{f(x|\theta_1)} \right)^2 \mathcal{V}''(1) + R_3(x), \quad (\text{A.1})$$



where  $\delta = (\theta_0 - \theta_1)$ , and  $R_3(x) = R_3(x; \theta_0, \theta_1)$  is the standard Taylor expansion remainder term. If we set  $\theta_1 = \theta_{\text{ob}}$ , then  $\mathcal{I}_V^T(\xi_1; x_{\text{ob}})$  becomes

$$\frac{1}{2}(\theta_0 - \theta_{\text{ob}})^2 I_{\text{ob}} \mathcal{V}'(1) + O((\theta_0 - \theta_{\text{ob}})^3). \quad (\text{A.2})$$

Next, for a sequence  $\{\theta^{(m)}\}$  such that  $|\theta^{(m)} - \theta_1| \leq \frac{1}{m}$ , we assume that the sequence  $\{R_3(\{x_{\text{ob}}, X_{\text{mis}}\}; \theta^{(m)}, \theta_1)\}$  is uniformly integrable. Then, inserting  $(X_{\text{mis}}, x_{\text{ob}})$  for  $x$  in (A.1), setting  $\theta_1 = \theta_{\text{ob}}$ , and taking an expectation with respect to  $f(X_{\text{mis}}|x_{\text{ob}}, \theta_{\text{ob}})$  we obtain

$$\mathcal{V}(1) - \frac{1}{2}(\theta_0 - \theta_{\text{ob}})^2 (I_{\text{ob}} \mathcal{V}'(1) - I_{\text{mis}} \mathcal{V}''(1)) + O((\theta_0 - \theta_{\text{ob}})^3). \quad (\text{A.3})$$

Hence, for  $\theta_1 = \theta_{\text{ob}}$ , we have

$$\mathcal{F}\mathcal{I}_V^T(\xi_2|\xi_1; x_{\text{ob}}) = \frac{I_{\text{ob}} \mathcal{V}'(1) + O(\theta_0 - \theta_{\text{ob}})}{I_{\text{ob}} \mathcal{V}'(1) - I_{\text{mis}} \mathcal{V}''(1) + O(\theta_0 - \theta_{\text{ob}})}, \quad (\text{A.4})$$

and letting  $\theta_{\text{ob}} \rightarrow \theta_0$  the result (1.41) follows.

## A.2 Split and combine proposals in reversible jump MCMC

The purpose of this appendix is to detail our implementation of split-combine moves in the BASCS code. We assume the reader is familiar with MCMC and RJMCMC algorithms. Those unfamiliar with MCMC we refer to [Gelman et al.](#)

(2013) and the appendix of Xu et al. (2014). Those unfamiliar with RJMCMC we refer to Richardson and Green (1997) and Green (1995). The basic properties of the algorithm follow from the reversibility condition and the theory of Markov chain convergence dealt with in many probability and stochastic processes books, for example Feller (1968).

We concentrate on the split proposals used in BASCS because they are more complex than the combine proposals. In particular, we detail the steps of a split proposal in BASCS for the extended full model (the most complex case considered). The corresponding combine proposals are straightforwardly obtained by solving the equations appearing in our split proposal scheme for the parameters of the combined source (i.e., the parameters of the yet to split source). Conditions that are required of newly split sources must also be satisfied when sources are combined. Following the algorithm is a short description of the reasons that its novel features are necessary in the current context.

Let  $\mu_j = (\mu_{jx}, \mu_{jy})$  be the location of the source the algorithm is attempting to split. Throughout this appendix, the parameters for the two newly proposed sources formed by a split will be subscripted as in the main parts of the paper except that a 1 will appear after the subscript  $j$  to indicate the first newly proposed source, and similarly a 2 will indicate the second newly proposed source e.g.  $\mu_{j1x}$  will denote the  $x$ -coordinate of the first newly proposed source formed by a split. The newly proposed sources are ordered so that  $\min(\gamma_{j11}, \gamma_{j12}) \leq \min(\gamma_{j21}, \gamma_{j22})$ , i.e., the smallest *gamma* distribution mean of the spectral model for the first newly proposed source is smaller than that of the second newly proposed source.

For the full model the ordering used is  $\gamma_{j1} \leq \gamma_{j2}$ , and for the spatial-only model it is  $\mu_{j1x} \leq \mu_{j2x}$ . These orderings are solely for the purposes of proposals; the label switching problem is discussed separately in Appendix B. A split proposal is performed as follows:

**Step 1:** Spectral parameters proposal: simulate  $u \sim \text{Uniform}(0, 1)$ .

(a) If  $u > 0.5$ , simulate  $u_1 \sim \text{Beta}(2, 2)$ ,  $t, v_2, v_3 \sim \text{Uniform}(0, 1)$  and  $v_4, v_5 \sim \text{gamma}(5, 5)$ . For  $a = \pi_j/u_1$  and  $b = (\pi_j + u_1 - 1)/u_1$  define

$$f(u_1, \pi_j) = \begin{cases} a & \text{if } a < 1 \\ 1 + e^{\frac{10}{a} - 10} \log(a) & \text{otherwise,} \end{cases} \quad (\text{A.5})$$

$$g(u_1, \pi_j) = \begin{cases} b & \text{if } b > 0 \\ be^{10b} \log(b) & \text{otherwise.} \end{cases} \quad (\text{A.6})$$

Then set

$$\pi_{j1} = tg(u_1, \pi_j) + (1 - t)h(u_1, \pi_j) \quad (\text{A.7})$$

$$\pi_{j2} = \frac{\pi_j - u_1\pi_{j1}}{1 - u_1} \quad (\text{A.8})$$

$$\gamma_{j11} = v_2\gamma_{j1} \quad (\text{A.9})$$

$$\gamma_{j21} = \frac{1 - v_{11}v_2}{1 - v_{11}}\gamma_{j1} \quad (\text{A.10})$$

$$\gamma_{j12} = \gamma_{j11} + \frac{v_3}{v_{12}}(\gamma_{j2} - \gamma_{j11}) \quad (\text{A.11})$$

$$\gamma_{j22} = \gamma_{j11} + \frac{1 - v_3}{1 - v_{12}}(\gamma_{j2} - \gamma_{j11}) \quad (\text{A.12})$$

$$\alpha_{j11} = v_4 \alpha_{j1} \quad (\text{A.13})$$

$$\alpha_{j12} = v_5 \alpha_{j2} \quad (\text{A.14})$$

$$\alpha_{j2l} = \frac{\rho_{j2l} \gamma_{j2l}^2}{A_{jl}} \quad \text{for } l = 1, 2, \quad (\text{A.15})$$

where

$$A_{jl} = \rho_{j1} \gamma_{jl}^2 \left(1 + \frac{1}{\alpha_{jl}}\right) - \rho_{j1l} \gamma_{j1l}^2 \left(1 + \frac{1}{\alpha_{j1l}}\right) - \rho_{j2l} \gamma_{j2l}^2, \quad (\text{A.16})$$

and  $\rho_{j1} = w_j \pi_j$ ,  $\rho_{j11} = w_{j1} \pi_{j1}$ ,  $\rho_{j12} = w_{j2} \pi_{j2}$ ,  $\rho_{j2} = w_j(1 - \pi_j)$ ,  $\rho_{j12} = w_{j1}(1 - \pi_{j1})$ , and  $\rho_{j22} = w_{j2}(1 - \pi_{j2})$ .

The split proposal is immediately rejected if  $\pi_{j1}$  is not between  $\min\left(1, \frac{\pi_j}{u_1}\right)$  and  $\max\left(0, \frac{\pi_j + u_1 - 1}{u_1}\right)$ , or  $\gamma_{j21} > \gamma_{j22}$ , or any of  $\gamma_{j11}$ ,  $\gamma_{j21}$ ,  $\gamma_{j12}$ ,  $\gamma_{j22}$  are outside the range of the spectral data  $E$ .

- (b) If  $u \leq 0.5$ , simulate  $\pi_{j1}, \pi_{j2} \sim \text{Beta}(10, 1)$ ,  $v_2, v_3 \sim \text{Uniform}(0, 1)$ , and  $v_4, v_5 \sim \text{Beta}(1, 5)$ . Then set  $u_1 = \pi_j$ ,  $\gamma_{j11} = \gamma_{j1}$ ,  $\gamma_{j21} = \gamma_{j2}$ ,  $\alpha_{j11} = \alpha_{j1}$ ,  $\alpha_{j21} = \alpha_{j2}$ , and

$$\gamma_{j12} = \gamma_{j1} + v_2(E_{\max} - \gamma_{j1}) \quad (\text{A.17})$$

$$\gamma_{j22} = \gamma_{j2} + v_3(E_{\max} - \gamma_{j2}) \quad (\text{A.18})$$

$$\alpha_{j12} = 20v_4 \quad (\text{A.19})$$

$$\alpha_{j22} = 20v_5. \quad (\text{A.20})$$

**Step 2:** Spatial parameters proposal: simulate  $u_1 \sim \text{Beta}(2, 2)$ ,  $u_2 \sim S_2 \text{Beta}(2, 2)$ , and  $u_3 \sim S_3 \text{Beta}(2, 2)$  (where  $S_2$  and  $S_3$  are independent random signs) and set

$$w_{j1} = w_j u_1 \tag{A.21}$$

$$w_{j2} = w_j (1 - u_1) \tag{A.22}$$

$$\mu_{j11} = \mu_{jx} - u_2 \sigma \sqrt{\frac{w_{j2}}{w_{j1}}} \tag{A.23}$$

$$\mu_{j21} = \mu_{jx} + u_2 \sigma \sqrt{\frac{w_{j1}}{w_{j2}}} \tag{A.24}$$

$$\mu_{j12} = \mu_{jy} - u_3 \sigma \sqrt{\frac{w_{j2}}{w_{j1}}} \tag{A.25}$$

$$\mu_{j22} = \mu_{jy} + u_3 \sigma \sqrt{\frac{w_{j1}}{w_{j2}}}. \tag{A.26}$$

In our algorithm  $\sigma = 1$  (tuning parameter).

**Step 3:** If  $(\mu_{j1x} - \mu_{j'x})^2 + (\mu_{j1y} - \mu_{j'y})^2 < (\mu_{j1x} - \mu_{j2x})^2 + (\mu_{j1y} - \mu_{j2y})^2$ , for some  $j' \in \{1, \dots, K\} \setminus \{j\}$ , then the split proposal is rejected. We also reject the split proposal if the proposed source locations are outside the convex hull of the spatial data  $(x, y)$ .

**Step 4:** To update  $s$  to  $s'$  randomly assign photon  $i$  to the first newly proposed source with probability  $p_i = p_{i1}/(p_{i1} + p_{i2})$ , and otherwise to the second newly proposed source, for each  $i \in \mathcal{I}_j$ . Here

$$p_{il} = w_{jl} f_{(\mu_{j11}, \mu_{j12})}(x_i, y_i) \sum_{r=1}^2 \pi_{jlr} g_{\alpha_{s_{jlr}}, \gamma_{s_{jlr}}}(E_i), \tag{A.27}$$

for  $l = 1, 2$ . We denote the probability of the particular allocation realized by  $P_{alloc}$ .

**Step 5:** Simulate  $u_{split} \sim \text{Uniform}(0, 1)$  and accept the proposed split if  $u_{split} < \min\{1, A\}$  where

$$A = \begin{cases} \frac{p(\Theta'_{K+1}, K+1, s' | x, y, E)}{p(\Theta_K, K, s | x, y, E)} \frac{d_{K+1}}{b_K P_{alloc}} \\ \times \frac{1}{\frac{1}{4} b_{2,2}(u_1) b_{2,2}(|u_2|) b_{2,2}(|u_3|)} \\ \times \frac{1}{g_{5,5}(v_4) g_{5,5}(v_5)} |J_a| & \text{if } u > 0.5 \text{ (Step 1)} \\ \\ \frac{p(\Theta'_K, K+1, s' | x, y, E)}{p(\Theta_K, K, s | x, y, E)} \frac{d_{K+1}}{b_K P_{alloc}} \\ \times \frac{1}{\frac{1}{4} b_{2,2}(|u_2|) b_{2,2}(|u_3|) b_{10,1}(\pi_{j1}) b_{10,1}(\pi_{j2})} \\ \times \frac{1}{b_{1,5}(v_4) b_{1,5}(v_5)} |J_b| & \text{otherwise.} \end{cases} \quad (\text{A.28})$$

Here, the notation  $b_{S,R}$  and  $g_{S,R}$  denotes the Beta( $S, R$ ) and gamma( $S, R$ ) densities, respectively, and

$$b_K = \begin{cases} \frac{1}{K} & \text{if } K = 1 \\ \frac{1}{2} \frac{1}{K} & \text{otherwise,} \end{cases} \quad (\text{A.29})$$

$$d_{K+1} = \begin{cases} \frac{1}{K+1} & \text{if } \|(\mu_{j1x}, \mu_{j1y}) - (\mu_{j2x}, \mu_{j2y})\|_2 \\ & \leq \|(\mu_{j1x}, \mu_{j1y}) - (\mu_{j'x}, \mu_{j'y})\|_2, \\ & \forall j' \in \{1, \dots, K\} / \{j\} \\ \frac{1}{2} \frac{1}{K+1} & \text{otherwise.} \end{cases} \quad (\text{A.30})$$

The Jacobian  $|J_a|$  is the determinant of a  $16 \times 16$  block matrix. The determinant of the upper-left  $6 \times 6$  block is  $w_j \sigma^2 / (u_1(1 - u_1))$ , and this is multiplied by the determinant of the lower-right block which is calculated numerically. The Jacobian  $|J_b|$  is  $20^2 (E_{\max} - \gamma_{j1})(E_{\max} - \gamma_{j2}) w_j \sigma^2 / (u_1(1 - u_1))$ .

There are two features of BASCS that are not explicitly dealt with in standard approaches. The first is that the distributions we split and combine are themselves mixture distributions. The second is that BASCS randomly chooses from two proposal schemes for the spectral parameters in Step 1 because a single approach does not address all the possibilities. The approach in Step 1(a) splits each *gamma* distribution in the current source's spectral model into two, thus forming two new spectral models for the newly proposed sources. The key aspect of this approach is that the new spectral models are designed to both be similar to the original. This makes sense in a situation where two similar sources have been mistaken for one. The approach in Step 1(b) is designed to split one true source into two, with each newly proposed source accounting for one *gamma* component of the true spectral model. Thus, the two new source spectral models each typically have nearly all their weight on a single *gamma*, which is almost invariably the

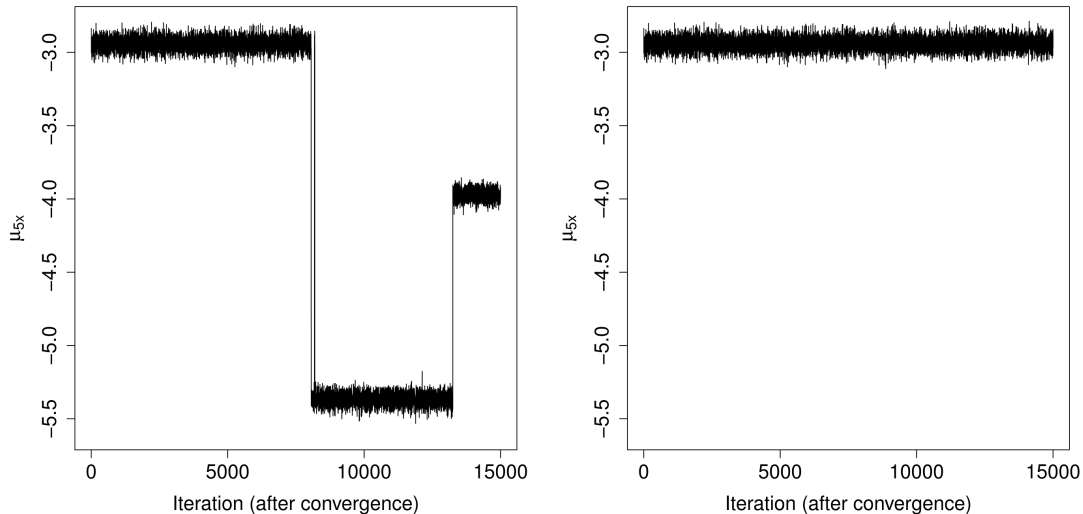
first component in the extended full spectral model (we sort the *gammas* by their means, in increasing order). Of course, we do not want to split a true source, but this split proposal is necessary in order to allow the reverse combine proposal, because the reversibility condition must be satisfied.

### A.3 Label switching

A computational challenge is that the enumeration, or labelling, of individual sources changes stochastically during the iterations of an RJMCMC algorithm (and even during the iterations of an MCMC algorithm for a mixture model with a known number of components). For example, Figure A.1 shows the value of  $\mu_{5x}$  at each iteration of our algorithm (after convergence) before and after the labelling has been corrected (the data are from the simulation study involving ten sources described in Section 2.5.1, and in particular,  $\mu_{5x}$  is the  $x$ -coordinate of the fifth source). Clearly, some such correction will be necessary in order for estimates such as that in (2.18) to be meaningful.

We implemented two approaches to relabelling and, in our real data analyses, they gave essentially identical results. The first method was to impose a hard constraint. In one dimension a hard constraint typically involves ordering the component locations, but it is not clear how best to impose such a constraint in two dimensions. As most of the source positions were precisely fitted, we simply ran the RJMCMC algorithm until convergence and then selected a posterior draw of the positions and weights to use as a reference. Running the algorithm again

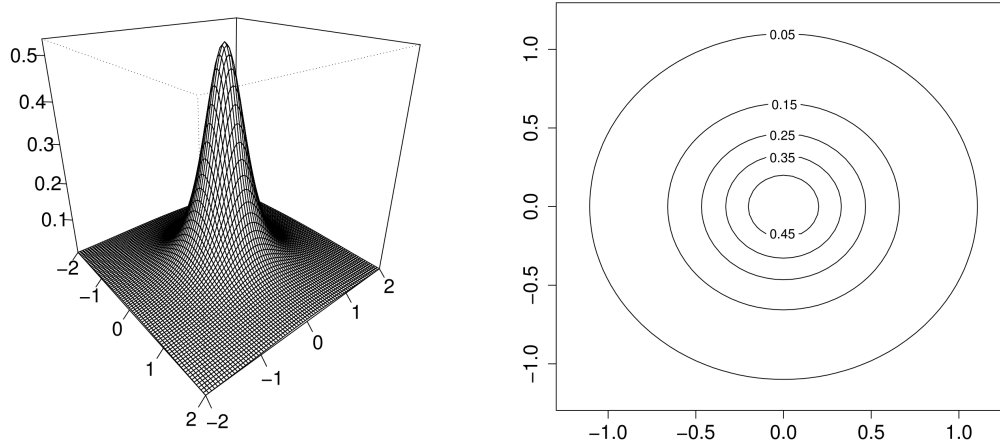




**Figure A.1:** Trace plot of the parameter  $\mu_{5x}$  from a simulation with ten sources (Section 2.5.1) before (left) and after (right) relabelling.

(or continuing the initial run), at each iteration we labelled the current source closest to the brightest reference source as source one, then we looked for the source closest to the second brightest reference source, and so on. As in the one dimensional case, this approach has the limitation that artificial ‘boundary’ effects may be introduced when the posteriors of two source positions overlap. These effects indicate that the real posterior uncertainty has not been correctly recovered (unless there is some real information to support a hard constraint in our prior). However, in our real data analyses there was no evidence of such boundaries because, for probable values of  $K$ , all the source positions were precisely fitted and there was little overlap between the posteriors of source positions. In the case of the *Chandra* observation and  $K = 14$ , the fact that the posteriors of the source locations are non-overlapping can be seen from Figure 2.11.

We also implemented the approach suggested by [Cron and West \(2011\)](#), by



**Figure A.2:** 2-D King profile density (left), and its contours (right).

modifying their publicly available code to work for our model. This method also uses a reference and is based on a loss function. At iteration  $t$ , the most likely assignment of each photon is computed treating the current parameter values as the true parameters, and then again treating the reference parameters as the true parameters. If, assuming the current parameter values, photon  $i$  is most likely to have originated from component two, but another origin is most likely when assuming the reference parameter values, then we say there is a mismatch in allocation of photon  $i$ . The method used by [Cron and West \(2011\)](#) is to choose the relabeling that minimizes the number of mismatches at iteration  $t$ , and then proceed to the next iteration. This second approach is substantially more computationally expensive than the first. Therefore we use the first approach online and apply the second only if there are potential ‘boundary’ effects (neither method is effected by the initial labels and therefore no problems are caused by applying

both).

## A.4 King profile

The functional form of the 2-D King profile is

$$f(d) = \frac{C}{(1 + (d/d_0)^2)^\eta} \quad (\text{A.31})$$

where

$$d(x, y, \omega) = \quad (\text{A.32})$$

$$\sqrt{(x \cos \omega + y \sin \omega)^2 + \frac{(y \cos \omega - x \sin \omega)^2}{(1 - \epsilon)^2}}. \quad (\text{A.33})$$

The constant  $C$  is determined numerically. The particular parameters we use for the 2-D King profile are as follows; off-axis angle  $\theta = 0$  arcmin, core radius  $d_0 = 0.6$  arcsec, power-law slope  $\eta = 1.5$ , ellipticity  $\epsilon = 0.00574$ . The resulting probability density is displayed in Figure A.2.

## A.5 Equivalence of the ML and warp bridge sampling estimators

Let

$$a_i = q_1(H_1^{-1}(H_2(x_i)))|J_{H_1^{-1}}(H_2(x_i))|\alpha(H_2(x_i)) \quad (\text{A.34})$$

$$b_i = q_2(H_2^{-1}(H_1(x_i)))|J_{H_2^{-1}}(H_1(x_i))|\alpha(H_1(x_i)) \quad (\text{A.35})$$

for  $i = 1, \dots, n$ . Equation (3.31) can then be rearranged to give

$$\frac{n_1}{n_2} \sum_{i=n_1+1}^n a_i - \hat{r} \sum_{i=1}^{n_1} b_i = 0 \quad (\text{A.36})$$

$$\implies 1 - \hat{r} \left( \sum_{i=n_1+1}^n q_2(x_i)|J_{H_2^{-1}}(H_2(x_i))|\alpha(H_2(x_i)) + \sum_{i=1}^{n_1} b_i \right) = 0 \quad (\text{A.37})$$

Setting  $T(x) = H_2^{-1}(H_1(x))$  we have  $|J_1(x)| = |J_{H_1^{-1}}(x)|/|J_{H_2^{-1}}(x)|$  and  $|J_2(x)| = 1/|J_1(x)|$ . It follows that the left hand side of (A.37) is equal to

$$1 - \hat{r} \sum_{i=1}^n \frac{1}{n_1 l_i + n_2 \hat{r}} = 0, \quad (\text{A.38})$$

where

$$l_i = \begin{cases} \frac{q_1(x_i, y_i)}{q_2(g(x_i, y_i))|J_2(x_i)|} & \text{for } i = 1, \dots, n_1 \\ \frac{q_1(g(x_i, y_i))|J_1(x_i)|}{q_2(x_i, y_i)} & \text{for } i = n_1 + 1, \dots, n. \end{cases} \quad (\text{A.39})$$

Utilizing the constraints in (3.19), we see that (A.38) is equivalent to

$$\hat{r} = \frac{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{\frac{n_1}{n} l_i + \frac{n_2}{n} \hat{r}}}{\frac{1}{n} \sum_{i=1}^n \frac{1}{\frac{n_1}{n} l_i + \frac{n_2}{n} \hat{r}}}, \quad (\text{A.40})$$

which is the definition of the group invariant sub-model MLE of  $r$  given in (3.28) (here we use the generalized group averaging specified in (3.30)), and therefore the proof is complete.

# References

- S. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.
- C. H. Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.
- J. O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *The Annals of Statistics*, pages 905–938, 2009.
- D. Blackwell. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 1, pages 93–102, 1951.
- D. M. Borth. A total entropy criterion for the dual problem of model discrimination and parameter estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 77–87, 1975.
- G. E. Box and W. J. Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.
- B. J. Brewer and D. Stello. Gaussian process modelling of asteroseismic data. *Monthly Notices of the Royal Astronomical Society*, 395(4):2226–2233, 2009.
- B. J. Brewer, D. Foreman-Mackey, and D. W. Hogg. Probabilistic catalogs for crowded stellar fields. *The Astronomical Journal*, 146(1):7, 2013.
- S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.
- P. S. Broos, L. K. Townsley, E. D. Feigelson, K. V. Getman, F. E. Bauer, and G. P. Garmire. Innovations in the analysis of chandra-acis observations. *The Astrophysical Journal*, 714(2):1582, 2010.
- K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

- A. J. Cron and M. West. Efficient classification-based relabeling in mixture models. *The American Statistician*, 65(1):16–20, 2011.
- Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitiit von markoffschen ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Science*, 8:85–108, 1963.
- J. E. Davis. Event pileup in charge-coupled devices. *The Astrophysical Journal*, 562(1):575, 2001.
- M. H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, pages 404–419, 1962.
- P. Dubath, L. Rimoldini, M. Süveges, J. Blomme, M. López, L. Sarro, J. De Ridder, J. Cuypers, L. Guy, I. Lecoœur, et al. Random forest automated supervised classification of hipparcos periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, 414(3):2602–2617, 2011.
- B. Efron and C. Morris. Data analysis using stein’s estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.
- V. V. Fedorov. *Theory of optimal experiments*. Elsevier, 1972.
- W. Feller. An introduction to probability theory, vol. i, vol. ii, 1968.
- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.
- K. Getman, E. Flaccomio, P. Broos, N. Grosso, M. Tsujimoto, L. Townsley, G. Garmire, J. Kastner, J. Li, F. Harnden Jr, et al. Chandra orion ultradeep project: observations and source lists. 160:319, 2005.
- J. Ginebra. On the measure of the information in a statistical experiment. *Bayesian Analysis*, 2(1):167–211, 2007.
- P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- J. Jasche and B. D. Wandelt. Methods for bayesian power spectrum inference with galaxy surveys. *The Astrophysical Journal*, 779(1):15, 2013.

- V. Kashyap, G. Micela, S. Sciortino, J. Harnden, F. R., and R. Rosner. 313:239, 1994.
- I. King. The structure of star clusters. i. an empirical density law. *The Astronomical Journal*, 67:471, 1962.
- M. L. Knoetig. Signal discovery, limits, and uncertainties with sparse on/off measurements: an objective bayesian analysis. *The Astrophysical Journal*, 790(2):106, 2014.
- A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan. A theory of statistical models for monte carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):585–604, 2003.
- A. Kong, P. McCullagh, X.-L. Meng, and D. L. Nicolae. Further explorations of likelihood theory for monte carlo integration. 2006.
- R. P. Kraft, D. N. Burrows, and J. A. Nousek. Determination of confidence limits for experiments with low numbers of counts. *The Astrophysical Journal*, 374:344–355, 1991.
- S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86, 1951.
- E. Laird, K. Nandra, A. Georgakakis, J. Aird, P. Barmby, C. Conselice, A. Coil, M. Davis, S. Faber, G. Fazio, et al. Aegis-x: the chandra deep survey of the extended groth strip. *The Astrophysical Journal Supplement Series*, 180(1):102, 2009.
- L. Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, pages 1419–1455, 1964.
- H. Lee, V. Kashyap, D. van Dyk, A. Connors, J. Drake, R. Izem, X.-L. Meng, S. Min, T. Park, P. Ratzlaff, A. Siemiginowska, and A. Zezas. Accounting for calibration uncertainties in x-ray analysis: Effective areas in spectral fitting. *Astrophysical Journal*, 731:126, 2011.
- D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.
- J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer, 2008.



- T. J. Loredo. Promise of bayesian inference for astrophysics. In *Statistical Challenges in Modern Astronomy*, pages 275–297. Springer, 1992.
- P. McCullagh. Quotient spaces and statistical models. *Canadian Journal of Statistics*, 27(3):447–456, 1999.
- G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.
- X.-L. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):pp. 552–586, 2002.
- X.-L. Meng and D. van Dyk. Minimum information ratio and relative augmentation function. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, pages 73–78, 1996.
- X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.
- N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- S. Mukherjee, E. D. Feigelson, G. J. Babu, F. Murtagh, C. Fraley, and A. Raftery. Three types of gamma-ray bursts. *The Astrophysical Journal*, 508(1):314, 1998.
- D. L. Nicolae and A. Kong. Measuring the relative information in allele-sharing linkage studies. *Biometrics*, 60(2):368–375, 2004.
- D. L. Nicolae, X.-L. Meng, and A. Kong. Quantifying the fraction of missing information for hypothesis testing in statistical and genetic studies. *Statistical Science*, 23(3):pp. 287–312, 2008.
- T. Park, V. L. Kashyap, A. Siemiginowska, D. A. Van Dyk, A. Zezas, C. Heinke, and B. J. Wargelin. Bayesian estimation of hardness ratios: Modeling and computations. *The Astrophysical Journal*, 652(1):610, 2006.
- T. Park, D. A. van Dyk, and A. Siemiginowska. Searching for narrow emission lines in X-ray spectra: Computation and methods. *The Astrophysical Journal*, 688:807–825, 2008.
- F. Primini and V. Kashyap. Determining x-ray source intensity and confidence bounds in crowded fields. *The Astrophysical Journal*, 796(1):24, 2014.

- A. Read and R. Saxton. 2-d psf parametrisation. Technical Report XMM-CCF-REL-280, 2012.
- A. Read, R. Saxton, M. Guainazzi, S. Rosen, and M. Stuhlinger. 2-d psf parametrisation. Technical Report XMM-CCF-REL-263, 2010.
- M. Reimherr, X.-L. Meng, and D. L. Nicolae. Being an informed bayesian: Assessing prior informativeness and prior likelihood conflict. *arXiv preprint arXiv:1406.5958*, 2014.
- J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard. On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1):10, 2011.
- S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.
- M. Safarzadeh, H. C. Ferguson, Y. Lu, H. Inami, and R. S. Somerville. Deconfusing blended field images using graphs and bayesian priors. *arXiv preprint arXiv:1408.2227*, 2014.
- T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, pages 1139–1154, 1993.
- C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- Q.-M. Shao and J. G. Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Series in Statistics, New York, 2000.
- S. Sherman. On a theorem of hardy, littlewood, polya, and blackwell. *Proceedings of the National Academy of Sciences*, 37(12):826–831, 1951.
- C. Stein. Notes on the comparison of experiments. *University of Chicago*, 1951.
- D. M. Titterington, A. F. Smith, U. E. Makov, et al. *Statistical analysis of finite mixture distributions*, volume 7. Wiley New York, 1985.
- B. Toman. Bayesian experimental design for multiple hypothesis testing. *Journal of the American Statistical Association*, 91(433):185–190, 1996.

- R. Umstätter, N. Christensen, M. Hendry, R. Meyer, V. Simha, J. Veitch, S. Vigeland, and G. Woan. Bayesian modeling of source confusion in lisa data. *Physical Review D*, 72(2):022001, 2005.
- D. A. van Dyk, A. Connors, V. Kashyap, and A. Siemiginowska. Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *The Astrophysical Journal*, 548:224–243, 2001.
- Y. Vardi. Empirical distributions in selection bias models. *The Annals of Statistics*, pages 178–203, 1985.
- A. F. Voter. A monte carlo method for determining free-energy differences and transition state theory rate constants. *The Journal of chemical physics*, 82:1890, 1985.
- J. Walmswell, J. Eldridge, B. Brewer, and C. Tout. A transdimensional bayesian method to infer the star formation history of resolved stellar populations. *Monthly Notices of the Royal Astronomical Society*, 435(3):2171–2186, 2013.
- M. C. Weisskopf, K. Wu, V. Trimble, S. L. O’Dell, R. F. Elsner, V. E. Zavlin, and C. Kouveliotou. A chandra search for coronal x-rays from the cool white dwarf gd 356. *The Astrophysical Journal*, 657(2):1026, 2007.
- M. Wiper, D. R. Insua, and F. Ruggeri. Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, 10(3), 2001.
- J. Xu, D. A. van Dyk, V. L. Kashyap, A. Siemiginowska, A. Connors, J. Drake, X.-L. Meng, P. Ratzlaff, and Y. Yu. A fully bayesian method for jointly fitting instrumental calibration and x-ray spectral models. *The Astrophysical Journal*, 794(2):97, 2014.