

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

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

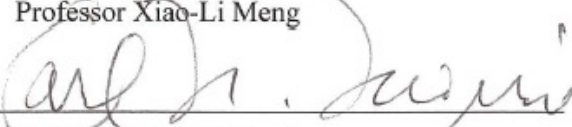
Topics in Bayesian Hierarchical Modeling and its Monte Carlo Computations

presented by **Hyung Suk Tak**

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 Carl N. Morris

Signature  _____

Typed name: Professor David A. van Dyk

Date: April 15, 2016

Topics in Bayesian Hierarchical Modeling and its Monte Carlo Computations

A dissertation presented

by

Hyung Suk Tak

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

April 2016

© 2016 Hyung Suk Tak

All rights reserved.

Topics in Bayesian Hierarchical Modeling and its Monte Carlo Computations

Abstract

The first chapter addresses a Beta-Binomial-Logit model that is a Beta-Binomial conjugate hierarchical model with covariate information incorporated via a logistic regression. Various researchers in the literature have unknowingly used improper posterior distributions or have given incorrect statements about posterior propriety because checking posterior propriety can be challenging due to the complicated functional form of a Beta-Binomial-Logit model. We derive data-dependent necessary and sufficient conditions for posterior propriety within a class of hyper-prior distributions that encompass those used in previous studies. Frequency coverage properties of several hyper-prior distributions are also investigated to see when and whether Bayesian interval estimates of random effects meet their nominal confidence levels.

The second chapter deals with a time delay estimation problem in astrophysics. When the gravitational field of an intervening galaxy between a quasar and the Earth is strong enough to split light into two or more images, the time delay is defined as the difference between their travel times. The time delay can be used to constrain cosmological parameters and can be inferred from the time series of brightness data of each image. To estimate the time delay, we construct a Gaussian hierarchical model based on a state-space representation for irregularly observed time series generated by a latent continuous-time Ornstein-Uhlenbeck process. Our Bayesian approach jointly infers model parameters via a Gibbs sampler. We also introduce a profile likelihood of the time delay as an approximation of its marginal posterior distribution.

The last chapter specifies a repelling-attracting Metropolis algorithm, a new Markov chain Monte Carlo method to explore multi-modal distributions in a simple and fast

manner. This algorithm is essentially a Metropolis-Hastings algorithm with a proposal that consists of a downhill move in density that aims to make local modes repelling, followed by an uphill move in density that aims to make local modes attracting. The downhill move is achieved via a reciprocal Metropolis ratio so that the algorithm prefers downward movement. The uphill move does the opposite using the standard Metropolis ratio which prefers upward movement. This down-up movement in density increases the probability of a proposed move to a different mode.

Contents

| | |
|---|------------|
| Contents | v |
| Acknowledgements | vii |
| 1 Data-dependent Posterior Propriety and Frequency Coverage Evaluation of a Bayesian Beta-Binomial-Logit Model | 1 |
| 1.1 Introduction | 1 |
| 1.2 Inferential model | 5 |
| 1.3 Posterior propriety | 7 |
| 1.4 Acceptance-rejection method | 22 |
| 1.5 Frequency method checking | 25 |
| 1.6 Example: Data of 18 baseball players | 28 |
| 1.7 Conclusion | 33 |
| 2 Bayesian Estimates of Astronomical Time Delays between Gravitationally Lensed Stochastic Light Curves | 35 |
| 2.1 Introduction | 35 |
| 2.2 A fully Bayesian model for time delay estimation | 42 |
| 2.3 Metropolis-Hastings within Gibbs sampler | 48 |
| 2.4 Profile likelihood of the time delay | 54 |
| 2.5 Time delay estimation strategy and numerical illustrations | 55 |

| | | |
|----------|--|------------|
| 2.6 | Conclusion | 65 |
| 3 | A Repelling-Attracting Metropolis Algorithm for Multimodality | 67 |
| 3.1 | Introduction | 67 |
| 3.2 | A repelling-attracting Metropolis algorithm | 69 |
| 3.3 | Numerical illustrations | 75 |
| 3.4 | Conclusion | 86 |
| | Appendices | 86 |
| A | Proofs of Theorem, Lemma, and Corollary in Chapter 1 | 90 |
| A.1 | Proof of Lemma 1.3.1 | 90 |
| A.2 | Proof of Theorem 1.3.1 | 91 |
| A.3 | Proof of Corollary 1.3.1 | 93 |
| A.4 | Proof of Theorem 1.3.4 | 94 |
| A.5 | Proof of Theorem 1.3.5 | 96 |
| B | Details on conditional distributions for the Gibbs sampler, profile likelihood, and sensitivity analysis in Chapter 2 | 98 |
| B.1 | Conditional distributions of $\mathbf{X}(\cdot)$ | 98 |
| B.2 | The likelihood function of parameters. | 99 |
| B.3 | Metropolis-Hastings within Gibbs sampler | 100 |
| B.4 | Profile likelihood approximately proportional to the marginal posterior . . . | 101 |
| B.5 | Sensitivity analyses | 102 |
| | Bibliography | 106 |

Acknowledgements

Thank you to Carl N. Morris for your invitation to the world of statistical research and for your sparing countless time to discuss ideas on hierarchical modeling. Thank you to Xiao-Li Meng and David A. van Dyk for your expanding my horizon into Astro-statistics and Markov chain Monte Carlo methods, for keeping me on the right track despite my innumerable trials and errors, and for elaborating all my crude ideas to be noteworthy. I thank you all for being a role model as a good researcher, a good teacher, and a good mentor.

I thank Joseph Kelly not only for teaching me valuable computational skills but for collaborating on developing an R package, Rgbp, my first project at Harvard. I would like to thank Kaisey Mandel for collaborating on my second project in Astrophysics and for supporting me with his balanced astrophysical and statistical knowledge. I also thank Vinay Kashyap and Aneta Siemiginowska for their guidance and support on the second project.

I thank my classmates and officemates, in particular, Joseph Lee, Qiuyi Han, Xufei Wang, Jiannan Lu, Yang Chen, David Jones, Bambo Sosina, Ed Kao, Ludovis Stourm, and Shi Yu for their friendship and support. I also thank my Korean friends, especially, Sunghwan Moon, Hyunyong Noh, Eunjoo Park, Keeseon Nam, Sukeun Jeong, Hyunsung Park, Junhyun Lee, Hanung Kim, Sujin Kim, Dongwoo Lee, Jiho Choi, Seonmi Park, Sokhyo Jo, Soyoun Shim, Hyeyoung You, Inkeun Song, and Eun Lee.

Lastly, thank you to my mother Kyungae Lee, father Jungam Tak, and sisters Hyojin Tak and Hyosun Tak for their constant encouragement, unwavering love, and support throughout the years.

To my family.

Chapter 1

Data-dependent Posterior Propriety and Frequency Coverage Evaluation of a Bayesian Beta-Binomial-Logit Model

1.1 Introduction

Binomial data from several independent groups sometimes have more variability than the assumed Binomial distribution for each group's count data. To account for this extra-Binomial variability, called overdispersion, a Beta-Binomial (BB) model (Skellam, 1948) puts a conjugate Beta prior distribution on unknown success probabilities by treating them as random effects. A Beta-Binomial-Logit (BBL) model (Williams, 1982; Kahn and Raftery, 1996) is one way to incorporate covariate information into the BB model. The BBL model

has a two-level structure as follows: For each of k independent groups ($j = 1, 2, \dots, k$),

$$y_j | p_j \stackrel{\text{indep.}}{\sim} \text{Bin}(n_j, p_j), \quad (1.1)$$

$$p_j | r, \boldsymbol{\beta} \stackrel{\text{indep.}}{\sim} \text{Beta}(rp_j^E, rq_j^E), \quad (1.2)$$

$$p_j^E = 1 - q_j^E \equiv E(p_j | r, \boldsymbol{\beta}) = \frac{\exp(\mathbf{x}_j^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})} \quad (1.3)$$

where y_j is the number of successful outcomes out of n_j trials, a sufficient statistic for the random effect p_j , $p_j^E = 1 - q_j^E$ denotes the expected random effect, $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jm})^\top$ is a covariate vector of length m for group j , $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_m)^\top$ is a vector of m logistic regression coefficients, and r represents the amount of prior information on p_j^E , considering that the Beta prior distribution in (1.2) concentrates on p_j^E as r increases (Albert, 1988). We focus only on a logit link function in (1.3) because it is canonical and is well defined for both binary ($n_j = 1$) and aggregate ($n_j \geq 2$) data. When there is no covariate with an intercept term, i.e., $\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_1$, the conjugate Beta distribution in (1.2) is exchangeable, and the BBL model reduces to the BB model.

The goal of our two-level conjugate modeling is to estimate random effects (p_1, p_2, \dots, p_k) for a comparison between groups. For example, this model can be used to estimate unknown true batting averages (random effects) of baseball players for a comparison among players based on their numbers of hits and at-bats possibly with their covariate information. Biologists may be interested in unknown true tumor incidence rates in analyzing litter data composed of each litter's number of tumor-bearing animals out of total number of animals at risk (Tamura and Young, 1987). The unknown true mortality rates on myocardial infarction can be estimated based on the death rate data collected from several independent clinical studies via a meta analysis (Gelman et al., 2013).

A Bayesian approach to the BBL model needs a joint hyper-prior distribution of r and $\boldsymbol{\beta}$ that affects posterior propriety. Though a proper joint hyper-prior distribution guarantees posterior propriety, various researchers have used improper hyper-prior distributions hoping for minimal impact on the posterior inference. The articles of Albert (1988) and Daniels

(1999) use $dr/(1+r)^2$ as a hyper-prior probability density function (PDF) for r , and independently an improper flat hyper-prior PDF for $\boldsymbol{\beta}$, $d\boldsymbol{\beta}$. Chapter 5 of Gelman et al. (2013) suggests putting an improper hyper-prior PDF on r , $dr/r^{1.5}$, and independently a proper standard Logistic distribution on β_1 when $\boldsymbol{x}^\top \boldsymbol{\beta} = \beta_1$. (They use a different parameterization: $p_j | \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ and $d\alpha d\beta/(\alpha + \beta)^{2.5}$. Transforming $r = \alpha + \beta$ and $p^E = \alpha/(\alpha + \beta)$, we obtain $dp^E dr/r^{1.5}$.) However, the paper of Albert (1988) does not address posterior propriety, the proposition in Daniels (1999) incorrectly concludes that posterior propriety holds regardless of the data, and Chapter 5 of Gelman et al. (2013) specifies an incorrect condition for posterior propriety.

To illustrate with an overly simple example for data-dependent conditions for posterior propriety, we toss two biased coins twice each ($n_j = 2$ for $j = 1, 2$). Let y_j indicate the number of Heads for coin j , and assume a BB model with $\boldsymbol{x}^\top \boldsymbol{\beta} = \beta_1$. If we use any proper hyper-prior PDF for r together with an improper flat density on an intercept term β_1 independently, posterior propriety holds except when both coins land either all Heads ($y_1 = y_2 = 2$) or all Tails ($y_1 = y_2 = 0$) as shown by an X in the diagram. Here the notation O means that the resulting posterior is proper. See Section 1.3.4 for details.

| | | | |
|----------------------|---|---|---|
| $y_1 \backslash y_2$ | 0 | 1 | 2 |
| 0 | X | O | O |
| 1 | O | O | O |
| 2 | O | O | X |

Also, there is a hyper-prior PDF for r that always leads to an improper posterior distribution regardless of the data. The article of Kass and Steffey (1989) adopts an improper joint hyper-prior PDF, $d\boldsymbol{\beta} dr/r$, without addressing posterior propriety. The paper of Kahn and Raftery (1996) uses the same improper hyper-prior PDF for r , dr/r , which they show is a Jeffreys' prior, and independently a proper multivariate Gaussian hyper-prior PDF for $\boldsymbol{\beta}$, declaring posterior propriety without a proof. However, the hyper-prior PDF dr/r used in both articles always leads to an improper posterior regardless of the data.

Making an inference unknowingly based on an improper posterior distribution is dangerous because the improper posterior distribution is not a probability distribution, and thus Markov chain Monte Carlo methods may draw samples from a nonexistent probability distribution (Hobert and Casella, 1996). We derive data-dependent necessary and sufficient conditions for posterior propriety of a Bayesian BBL model equipped with various joint hyper-prior distributions, and summarize these conditions in Figure 1.1, the centerpiece of this article. We mainly work on a class of hyper-prior PDFs for r , $dr/(t+r)^{u+1}$, where t is non-negative and u is positive. It includes a proper $dr/(1+r)^2$ (Albert, 1988; Daniels, 1999) and an improper $dr/r^{1.5}$ (Gelman et al., 2013) as special cases. Independently the hyper-prior PDF for β that we consider is improper flat (Lebesgue measure) for its intended minimal impact on posterior inference or any proper one. When a posterior distribution is improper due to improper hyper-prior distributions, one possible alternative is to use proper hyper-prior distributions that can mimic the behavior of improper choices, e.g., $dr/(t+r)^{u+1}$ with a small constant t to mimic dr/r^{u+1} and a diffuse Gaussian distribution for β to mimic its improper flat choice.

We compare operating characteristics of several hyper-prior distributions for r via repeated sampling coverage simulations, which we call frequency method checking (Morris and Christiansen, 1997; Morris and Tang, 2011b). Here, the purpose of frequency method checking is to see when and whether the posterior intervals of the random effects p_j meet their nominal confidence levels. Because conditions for posterior propriety with specific improper hyper-prior distributions are data-dependent, we estimate the coverage rates based only on the simulated data sets that meet the conditions for posterior propriety.

This chapter is organized as follows. We derive an equivalent inferential model of the Bayesian BBL model in Section 1.2. We derive necessary and sufficient conditions for posterior propriety, address posterior propriety in past studies, discuss possible alternatives when posterior distributions are improper, and provide two simple examples related to the conditions for posterior propriety in Section 1.3. An acceptance-rejection method to fit a

Bayesian BBL model is specified in Section 1.4 and frequency method checking techniques are introduced in Section 2.2. We conduct frequency method checking to compare several hyper-prior distributions in 2.2.

1.2 Inferential model

One advantage of the BBL model is that it allows the shrinkage interpretation in inference (Kahn and Raftery, 1996). For $j = 1, 2, \dots, k$, the conditional posterior distribution of a random effect p_j given hyper-parameters and data is

$$p_j \mid r, \boldsymbol{\beta}, \mathbf{y} \stackrel{\text{indep.}}{\sim} \text{Beta}(rp_j^E + y_j, rq_j^E + (n_j - y_j)) \quad (1.4)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_k)^\top$. The posterior mean of the conditional posterior distribution in (1.4) is $\hat{p}_j \equiv (1 - B_j)\bar{y}_j + B_j p_j^E$. This mean is a convex combination of the observed proportion $\bar{y}_j = y_j/n_j$ and the expected random effect p_j^E weighted by the relative amount of information in the prior compared to the data, called a shrinkage factor $B_j = r/(r + n_j)$; r determines the precision of p_j^E and n_j determines the precision of \bar{y}_j . If the conjugate prior distribution contains more information than the observed data, i.e., ensemble sample size r exceeds individual sample size n_j , then the posterior mean shrinks more towards p_j^E than towards \bar{y}_j . The posterior variance of this conditional posterior distribution in (1.4) is a quadratic function of \hat{p}_j , i.e., $\hat{p}_j(1 - \hat{p}_j)/(r + n_j + 1)$.

The conjugate Beta prior distribution of random effects in (1.2) has unknown hyper-parameters, r and $\boldsymbol{\beta}$. Assuming r and $\boldsymbol{\beta}$ are independent a priori, we introduce their joint hyper-prior PDF as follows:

$$\pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) = f(r)g(\boldsymbol{\beta}) \propto \frac{g(\boldsymbol{\beta})}{(t+r)^{u+1}}, \text{ for } t \geq 0 \text{ and } u > 0. \quad (1.5)$$

This class of hyper-prior PDFs for r , i.e., $dr/(t+r)^{u+1}$, is proper if $t > 0$ and improper if $t = 0$. A hyper-prior PDF for a uniform shrinkage prior on r , transformed from a uniform prior on a shrinkage factor $dB = d\{r/(t+r)\}$, is $dr/(t+r)^2$ with $u = 1$ for any positive constant t (Morris

and Christiansen, 1997). This uniform shrinkage prior is known to have good frequentist properties for Bayesian estimates (Strawderman, 1971; Morris and Christiansen, 1997). A special case of the uniform shrinkage prior density function is $dr/(1+r)^2$ corresponding to $t = 1$ used by Albert (1988) and Daniels (1999). As t goes to zero, a proper uniform shrinkage prior density, $dr/(t+r)^2$, becomes close to an improper hyper-prior PDF dr/r^2 . This improper choice, dr/r^2 , is free of an arbitrary constant t and is the most conservative choice that leads to the widest posterior intervals for random effects compared to those obtained by any uniform shrinkage prior (Morris and Christiansen, 1997). Chapter 5 of Gelman et al. (2013) suggests using $dr/r^{1.5}$ as a diffuse hyper-prior PDF, which corresponds to $u = 0.5$ and $t = 0$, together with a standard Logistic distribution on $\boldsymbol{\beta}$. Jeffreys' prior dr/r (Kahn and Raftery, 1996) is not included in the class because it always leads to an improper posterior distribution regardless of the data¹; see Section 1.3.2. The hyper-prior PDF for $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, can be any proper PDF or an improper flat density.

The marginal distribution of the data follows independent Beta-Binomial distributions (Skellam, 1948) with random effects integrated out. The probability mass function for the Beta-Binomial distribution is, for $j = 1, 2, \dots, k$,

$$\pi_{\text{obs}}(y_j | r, \boldsymbol{\beta}) = \binom{n_j}{y_j} \frac{B(y_j + rp_j^E, n_j - y_j + rq_j^E)}{B(rp_j^E, rq_j^E)} \quad (1.6)$$

where the notation $B(a, b)$ indicates a beta function defined as $\int_0^1 v^{a-1}(1-v)^{b-1}dv$ for positive constants a and b . The probability mass function in (1.6) depends on $\boldsymbol{\beta}$ because the expected random effects, $\{p_1^E, p_2^E, \dots, p_k^E\}$, are a function of $\boldsymbol{\beta}$ as shown in (1.3). The likelihood function of r and $\boldsymbol{\beta}$ is the product of these Beta-Binomial probability mass functions being treated as expressions in r and $\boldsymbol{\beta}$, i.e.,

$$L(r, \boldsymbol{\beta}) = \prod_{j=1}^k \pi_{\text{obs}}(y_j | r, \boldsymbol{\beta}) = \prod_{j=1}^k \binom{n_j}{y_j} \frac{B(y_j + rp_j^E, n_j - y_j + rq_j^E)}{B(rp_j^E, rq_j^E)}. \quad (1.7)$$

¹If the symbol A represents a second-level variance component in a two-level Gaussian multilevel model, e.g., $y_j | \mu_j \sim \text{Normal}(\mu_j, 1)$ and $\mu_j | A \sim \text{Normal}(0, A)$, then A is proportional to $1/r$. The improper hyper-prior PDF $dr/r^2 = -d(1/r)$ corresponds to dA leading to Stein's harmonic prior (Morris and Tang, 2011b), $dr/r^{1.5}$ corresponds to dA/\sqrt{A} (Gelman et al., 2013), and dr/r is equivalent to an inappropriate choice dA/A (Morris and Lysy, 2012).

When $n_j = 1$, this likelihood function reduces to the one of a logistic regression model:

$$L(r, \boldsymbol{\beta}) = \prod_{j=1}^k (p_j^E)^{y_j} (1 - p_j^E)^{1-y_j} = \prod_{j=1}^k \left(\frac{\exp(\mathbf{x}_j^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})} \right)^{y_j} \left(\frac{1}{1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})} \right)^{1-y_j}, \quad (1.8)$$

which is free of r . Since the data tell nothing about r when $n_j = 1$ for all j , it is better not to make any inference on the random effects, p_1, p_2, \dots, p_k , via a Bayesian BBL model unless we have prior information on r .

The joint posterior density function of hyper-parameters, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$, is proportional to their likelihood function in (1.7) multiplied by the joint hyper-prior PDF in (1.5):

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y}) \propto \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \times L(r, \boldsymbol{\beta}). \quad (1.9)$$

Finally, the full posterior density function of $\mathbf{p} = (p_1, p_2, \dots, p_k)^\top$, r , and $\boldsymbol{\beta}$ is

$$\begin{aligned} \pi_{\text{full.post}}(\mathbf{p}, r, \boldsymbol{\beta} \mid \mathbf{y}) &\propto \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \times \prod_{j=1}^k \pi_{\text{obs}}(y_j \mid p_j) \times \pi_{\text{prior}}(p_j \mid r, \boldsymbol{\beta}) \\ &\propto \pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y}) \times \prod_{j=1}^k \pi_{\text{cond.post}}(p_j \mid r, \boldsymbol{\beta}, \mathbf{y}) \end{aligned} \quad (1.10)$$

where the distribution for the prior density function of random effect j , $\pi_{\text{prior}}(p_j \mid r, \boldsymbol{\beta})$, is specified in (1.2), and the distribution of the conditional posterior density of random effect j , $\pi_{\text{cond.post}}(p_j \mid r, \boldsymbol{\beta}, \mathbf{y})$, is specified in (1.4).

1.3 Posterior propriety

The full posterior density function in (1.10) is proper if and only if $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is proper because $\prod_{j=1}^k \pi_{\text{cond.post}}(p_j \mid r, \boldsymbol{\beta}, \mathbf{y})$ is a product of independent and proper Beta density functions. We therefore focus on posterior propriety of $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$.

Definition 1.3.1. *Group j whose observed number of successes is neither 0 nor n_j , i.e., $1 \leq y_j \leq n_j - 1$, is called an interior group. Similarly, group j is extreme if its observed number of successes is either 0 or n_j . The symbol $W_{\mathbf{y}}$ denotes the set of indices corresponding*

to interior groups, i.e., $W_y \subseteq \{1, 2, \dots, k\}$, and k_y is the number of interior groups, i.e., the number of indices in W_y . We use W_y^c to represent the set of $k - k_y$ indices for extreme groups. The notation $X \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k)^\top$ refers to the $k \times m$ covariate matrix of all groups ($k \geq m$) and X_y is the $k_y \times m$ covariate matrix of the interior groups.

The subscript y emphasizes the data-dependence of k_y , W_y , and X_y . The rank of X_y can be smaller than m when X is of full rank m because we obtain X_y by removing rows of extreme groups from X . If all groups are interior, then $k_y = k$ and $X_y = X$. If all groups are extreme, then $k_y = 0$ and X_y is not defined.

1.3.1 Conditions for posterior propriety

In Figure 1.1, we summarize the necessary and sufficient conditions for posterior propriety according to different hyper-prior PDFs, $f(r)$ and $g(\boldsymbol{\beta})$, under two settings: The data contain at least one interior group ($1 \leq k_y \leq k$) and the data contain only extreme groups ($k_y = 0$).

To prove these conditions, we divide the first setting ($1 \leq k_y \leq k$) into two: A setting where at least one interior group and at least one extreme group exist ($1 \leq k_y \leq k - 1$) and a setting where all groups are interior ($k_y = k$). The key to proving conditions for posterior propriety is to derive certain lower and upper bounds for $L(r, \boldsymbol{\beta})$ that factor into a function of r and a function of $\boldsymbol{\beta}$. We first derive lower and upper bounds for the Beta-Binomial probability mass function of group j with respect to r and $\boldsymbol{\beta}$ because $L(r, \boldsymbol{\beta})$ is just the product of these probability mass functions of all groups.

Lemma 1.3.1. *Lower and upper bounds for the Beta-Binomial probability mass function for interior group j with respect to r and $\boldsymbol{\beta}$ are $rp_j^E q_j^E / (1+r)^{n_j-1}$ and $rp_j^E q_j^E / (1+r)$, respectively, up to a constant multiple. Those for extreme group j with $y_j = n_j$ are $(p_j^E)^{n_j}$ and p_j^E , each, and those for extreme group j with $y_j = 0$ are $(q_j^E)^{n_j}$ and q_j^E , respectively, up to a constant multiple.*

Proof. See Appendix A.1. □

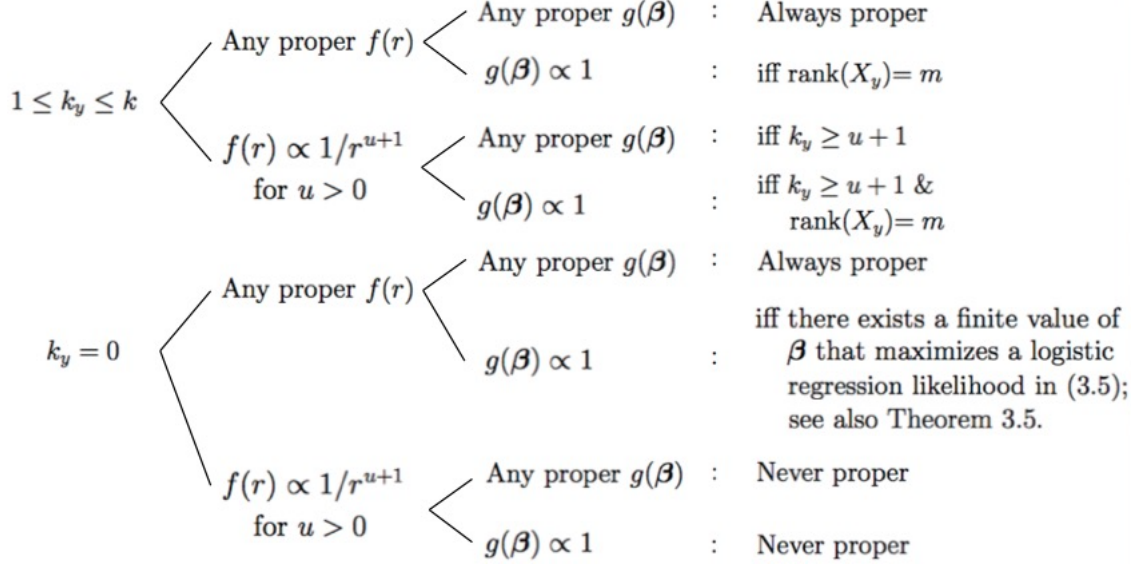


Figure 1.1: Necessary and sufficient conditions for posterior propriety of $\pi_{\text{hyp.post}}(r, \beta \mid \mathbf{y})$ according to $\pi_{\text{hyp.prior}}(r, \beta) = f(r)g(\beta)$ under two settings: The data contain at least one interior group ($1 \leq k_y \leq k$) and the data contain only extreme groups ($k_y = 0$). The condition, $\text{rank}(X_y) = m$, implicitly requires $k_y \geq m$ because X_y is a $k_y \times m$ matrix.

Lemma 1.3.1 shows that our bounds for the Beta-Binomial probability mass function for either interior or extreme group j with respect to r and β factor into a function of r and a function of β . Because $L(r, \beta)$ is a product of these Beta-Binomial probability mass functions of all groups, bounds for $L(r, \beta)$ also factor into a function of r and a function of β . Next we derive certain lower and upper bounds for $L(r, \beta)$ with respect to r and β under the first setting where all groups are interior.

Lemma 1.3.2. *When all groups are interior ($k_y = k$), $L(r, \beta)$ can be bounded by*

$$c_1 \frac{r^k \prod_{j=1}^k p_j^E q_j^E}{(1+r)^{\sum_{j=1}^k (n_j-1)}} \leq L(r, \beta) \leq c_2 \frac{r^k \prod_{j=1}^k p_j^E q_j^E}{(1+r)^k} \quad (1.11)$$

where c_1 and c_2 are constants that do not depend on r and β .

Proof. Without any extreme groups in the data, an upper bound for $L(r, \beta)$ is the product of the k upper bounds for the Beta-Binomial probability mass function of each interior group in (A.4), i.e., $r^k (\prod_{j=1}^k p_j^E q_j^E) / (1+r)^k$. Similarly, a lower bound for $L(r, \beta)$ is the product of the k lower bounds for the Beta-Binomial probability mass function of each interior group

in (A.7), i.e., $r^k(\prod_{j=1}^k p_j^E q_j^E)/(1+r)^{\sum_{j=1}^k (n_j-1)}$. It is clear that both bounds factor into a function of r and a function of $\boldsymbol{\beta}$. \square

When all groups are interior, the joint posterior density function $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ equipped with any joint hyper-prior PDF $\pi_{\text{hyp.prior}}(r, \boldsymbol{\beta})$ is proper if

$$\int_{\mathbf{R}^m} \int_0^\infty \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \times \frac{r^k \prod_{j=1}^k p_j^E q_j^E}{(1+r)^k} dr d\boldsymbol{\beta} < \infty \quad (1.12)$$

because $r^k \prod_{j=1}^k p_j^E q_j^E / (1+r)^k$ is the upper bound for $L(r, \boldsymbol{\beta})$ specified in (1.11). Also, the joint posterior density function $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is improper if

$$\int_{\mathbf{R}^m} \int_0^\infty \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \times \frac{r^k \prod_{j=1}^k p_j^E q_j^E}{(1+r)^{\sum_{j=1}^k (n_j-1)}} dr d\boldsymbol{\beta} = \infty \quad (1.13)$$

because $r^k \prod_{j=1}^k p_j^E q_j^E / (1+r)^{\sum_{j=1}^k (n_j-1)}$ is the lower bound for $L(r, \boldsymbol{\beta})$ in (1.11).

Theorem 1.3.1. *When all groups are interior ($k_y = k$), the joint posterior density function of hyper-parameters, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$, equipped with a proper hyper-prior density function on r , $f(r)$, and independently an improper flat hyper-prior density function on $\boldsymbol{\beta}$, $g(\boldsymbol{\beta}) \propto 1$, is proper if and only if $\text{rank}(X) = m$.*

Proof. See Appendix A.2. \square

The condition for posterior propriety with a proper hyper-prior PDF for r is the same as the condition for posterior propriety when r is a completely known constant due to the factorization of the bounds for $L(r, \boldsymbol{\beta})$ in (1.11). Thus, the condition for posterior propriety in Theorem 1.3.1 arises only from the improper hyper-prior PDF for $\boldsymbol{\beta}$.

Theorem 1.3.2. *When all groups are interior ($k_y = k$), the joint posterior density function of hyper-parameters, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$, equipped with $f(r) \propto 1/r^{u+1}$ for positive u and independently a proper hyper-prior density function on $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, is proper if and only if $k \geq u + 1$.*

Proof. The $\boldsymbol{\beta}$ part of the upper bound for $L(r, \boldsymbol{\beta})$ in Lemma 1.3.2, i.e., $\prod_{j=1}^k p_j^E q_j^E$, is always less than one. Thus, the upper bound for $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ up to a normalizing constant factors into a function of r and a function of $\boldsymbol{\beta}$ as follows:

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y}) \propto f(r)g(\boldsymbol{\beta})L(r, \boldsymbol{\beta}) < \frac{r^{k-(u+1)}g(\boldsymbol{\beta})}{(1+r)^k}. \quad (1.14)$$

The integration of this upper bound with respect to r is finite if $k \geq u + 1$ because in this case we can bound the r part by $1/(1+r)^{u+1}$ whose integration with respect to r is always finite. The integration of $g(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is finite because $g(\boldsymbol{\beta})$ is a proper probability density function.

If $k < u + 1$, then the integration of the lower bound for $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is not finite because there is r^k in the numerator of the lower bound for $L(r, \boldsymbol{\beta})$ in Lemma 1.3.2. Specifically, once multiplying $f(r)$ ($\propto dr/r^{u+1}$) by r^k , we know that $r^{k-(u+1)}$ goes to infinity as r approaches zero if $k < u + 1$. \square

The condition for posterior propriety when $\boldsymbol{\beta}$ has a proper hyper-prior distribution is the same as the condition for posterior propriety when $\boldsymbol{\beta}$ is not a parameter to be estimated ($m = 0$) due to the factorization of bounds for $L(r, \boldsymbol{\beta})$ in (1.11). Thus, the condition for posterior propriety arises solely from the improper hyper-prior PDF for r .

Theorem 1.3.3. *When all groups are interior ($k_y = k$), the joint posterior density function of hyper-parameters, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$, equipped with the joint hyper-prior density function $\pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \propto 1/r^{u+1}$ for positive u is proper if and only if (i) $k \geq u + 1$ and (ii) $\text{rank}(X) = m$.*

Proof. Based on the upper bound for $L(r, \boldsymbol{\beta})$ in Lemma 1.3.2, the upper bound for $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ up to a normalizing constant factors into a function of r and a function of $\boldsymbol{\beta}$ as follows:

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y}) \propto \pi_{\text{hyp.prior}}(r, \boldsymbol{\beta})L(r, \boldsymbol{\beta}) \leq \frac{r^{k-(u+1)}}{(r+1)^k} \prod_{j=1}^k p_j^E q_j^E. \quad (1.15)$$

The double integration on the upper bound in (1.15) with respect to r and $\boldsymbol{\beta}$ is finite if and only if (i) $k \geq u + 1$ for the r part as proved in Theorem 1.3.2 and (ii) the $k \times m$ covariate matrix of all groups X has a full rank m for the $\boldsymbol{\beta}$ part as proved in Theorem 1.3.1.

If at least one condition is not met, then $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ becomes improper as proved in Theorem 1.3.1 and 1.3.2. \square

The conditions for posterior propriety in Theorem 1.3.3 are the combination of the condition in Theorem 1.3.1 and that in Theorem 1.3.2 because of the factorization of bounds for $L(r, \boldsymbol{\beta})$.

We begin discussing the conditions for posterior propriety under the second setting with at least one interior group and at least one extreme group in the data ($1 \leq k_y \leq k - 1$).

Corollary 1.3.1. *With at least one interior group and at least one extreme group in the data ($1 \leq k_y \leq k - 1$), posterior propriety is determined solely by interior groups, not by extreme groups.*

Proof. See Appendix A.3. \square

Corollary 1.3.1 means that we can remove all the extreme groups from the data to determine posterior propriety, treating the remaining interior groups as a new data set ($k_y = k$). Then we can apply Theorem 1.3.1, 1.3.2, or 1.3.3 to the new data set. If posterior propriety holds with only the interior groups, then posterior propriety with the original data with the combined interior and extreme groups ($1 \leq k_y \leq k - 1$) also holds. Corollary 1.3.1 justifies combining the first and second settings as shown in Figure 1.1.

We start specifying the conditions for posterior propriety under the third setting where there are no interior groups in the data ($k_y = 0$).

Lemma 1.3.3. *When all groups are extreme ($k_y = 0$), $L(r, \boldsymbol{\beta})$ can be bounded by*

$$c_3 \prod_{j=1}^k (p_j^E)^{n_j \times I_{\{y_j=n_j\}}} (q_j^E)^{n_j \times I_{\{y_j=0\}}} \leq L(r, \boldsymbol{\beta}) \leq c_4 \prod_{j=1}^k (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \quad (1.16)$$

where c_3 and c_4 are constants that do not depend on r and $\boldsymbol{\beta}$, and $I_{\{D\}}$ is the indicator function of D .

Proof. A lower bound for the Beta-Binomial probability mass function of extreme group j is either $(p_j^E)^{n_j}$ in (A.9) or $(q_j^E)^{n_j}$ in (A.10) depending on whether $y_j = n_j$ or $y_j = 0$. Thus, the product of k lower bounds for the Beta-Binomial probability mass functions of extreme groups, i.e., $\prod_{j=1}^k (p_j^E)^{n_j \times I_{\{y_j=n_j\}}} (q_j^E)^{n_j \times I_{\{y_j=0\}}}$, bounds $L(r, \boldsymbol{\beta})$ from below.

The product of the k upper bounds for the Beta-Binomial probability mass functions of extreme groups in (A.8) or (A.10), i.e., $\prod_{j=1}^k (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}}$, bounds $L(r, \boldsymbol{\beta})$ from above. □

The upper and lower bounds in (1.16) are free of r , indicating that the hyper-prior distribution of r must be proper for posterior propriety in this case ($k_y = 0$). If the hyper-prior distribution of $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, is also proper, the resulting posterior is automatically proper and we do not need to check posterior propriety. However, the posterior can be improper when $g(\boldsymbol{\beta})$ is improper. The next theorem deals with a case where $g(\boldsymbol{\beta}) \propto 1$.

Theorem 1.3.4. *When all groups are extreme ($k_y = 0$), the posterior density function of hyper-parameters, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$, equipped with a proper hyper-prior density function for r , $f(r)$, and independently $g(\boldsymbol{\beta}_1) \propto 1$, is proper if and only if there exists a finite value of $\boldsymbol{\beta}$ that maximizes the upper bound in (1.16) up to a constant, i.e.,*

$$\prod_{j=1}^k \left(\frac{\exp(\mathbf{x}_j^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})} \right)^{I_{\{y_j=n_j\}}} \left(\frac{1}{1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})} \right)^{I_{\{y_j=0\}}} . \quad (1.17)$$

Proof. See Appendix A.4. □

The function in (1.17) is essentially the same as the likelihood function of a logistic regression in (1.8) because the powers in (1.17) are either one or zero with $I_{\{y_j=0\}} = 1 - I_{\{y_j=n_j\}}$. Thus, the value of $\boldsymbol{\beta}$ that maximizes (1.17) is the same as the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ in (1.8) for which we set $y_j = 1$ if $y_j \geq 1$. A quick way to check whether there exists a finite value of $\boldsymbol{\beta}$ that maximizes (1.17) is to fit a logistic regression model after

setting $y_j = 1$ if $y_j \geq 1$, using any statistical software, e.g., `glm` in R (R Development Core Team, 2016). If no errors emerge, then the finite MLE of $\boldsymbol{\beta}$ exists; its uniqueness is guaranteed if the MLE exists in a logistic regression (Jacobsen, 1989). However, Theorem 1.3.4 is inconvenient in that we need to fit a logistic regression model to check posterior propriety. The next theorem specifies more descriptive sufficient conditions for posterior propriety that do not require fitting a logistic regression, which are also necessary conditions when there is only an intercept term, $\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_1$ for all j .

Theorem 1.3.5. *When all groups are extreme ($k_y = 0$), the posterior density function of hyper-parameters, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$, equipped with a proper hyper-prior density function for r , $f(r)$, and independently $g(\beta_1) \propto 1$, is proper if (i) there are at least m clusters of groups whose covariate values are the same within each cluster and different between clusters, and (ii) in each cluster there are at least one group of all successes and at least one group of all failures. The $k \times m$ covariate matrix X is assumed to be of full rank m . These two conditions are also necessary conditions when $\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_1$.*

Proof. See Appendix A.5. □

When $\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_1$, the necessary and sufficient conditions in Theorem 1.3.5 simply reduce to having at least one group with all successes and at least one group with all failures in the data. Theorem 4 of Natarajan and Kass (2000) shows that this reduced condition is the same as the condition in Theorem 1.3.4, i.e., there exists a finite value of $\boldsymbol{\beta}$ that maximizes (1.17).

The two conditions in Theorem 1.3.5 are only sufficient conditions when there are covariates. For necessary conditions in this case, we need to show that integration of the lower bound in (1.16) with respect to $\boldsymbol{\beta}$ is not finite when either conditions in Theorem 1.3.5 are not met. However, the integration itself seems mathematically intractable. If either conditions in Theorem 1.3.5 are not met, we need to go back to Theorem 1.3.4, checking the existence of a finite value of $\boldsymbol{\beta}$ that maximizes (1.17) by fitting a logistic regression.

Theorem 1.3.6. *When all groups are extreme ($k_y = 0$), the posterior density function of hyper-parameters $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$, equipped with any improper hyper-prior density function $f(r)$ and independently any hyper-prior density $g(\boldsymbol{\beta})$, is always improper.*

Proof. Because the lower bound for $L(r, \boldsymbol{\beta})$ in Lemma 1.3.3 is free of r , $L(r, \boldsymbol{\beta})$ cannot make the integration of $f(r)$ finite when $f(r)$ is improper. Thus, $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ should always be improper under this setting. \square

1.3.2 Posterior propriety in previous studies

The article of Albert (1988) does not address posterior propriety for $d\boldsymbol{\beta}dr/(1+r)^2$. Our work shows that the condition for posterior propriety when $1 \leq k_y \leq k$ is $\text{rank}(X_y) = m$, i.e., the covariate matrix of interior groups is of full rank m . However, when $k_y = 0$, posterior propriety is unknown except for a case where only an intercept term is used ($\mathbf{x}^\top \boldsymbol{\beta} = \beta_1$), see Figure 1.1.

The proposition (1c to be specific) in Daniels (1999) for posterior propriety of the Bayesian BBL model with the same hyper-prior PDF as Albert (1988) argues that the posterior distribution is always proper. However, its proof is based on a limited case with only an intercept term, $\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_1$. Under this simplified setting, if there is only one extreme group with two trials ($y_1 = 2, n_1 = 2$), the resulting joint posterior density function of r and β_1 is

$$\pi_{\text{hyp.post}}(r, \beta_1 \mid \mathbf{y}) \propto \frac{(1 + rp^E)p^E}{(1 + r)^3}. \quad (1.18)$$

The integration of (1.18) with respect to β_1 is not finite because $p^E = \exp(\beta_1)/(1 + \exp(\beta_1))$ converges to one as β_1 approaches infinity. Figure 1.1 shows that at least one interior group is required in the data for posterior propriety of the Bayesian BBL model under the simplified setting ($\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_1$) of Daniels (1999). Moreover, if all groups are extreme in the data under the simplified setting with an intercept term, the posterior is proper if and only if there exist at least one extreme group with all successes ($\sum_{j=1}^k I_{\{y_j=n_j\}} \geq 1$) and one extreme group with all failures ($\sum_{j=1}^k I_{\{y_j=0\}} \geq 1$) as shown in Figure 1.1. In our counter-example, there

is only one extreme group with all successes, and thus the resulting posterior in (1.18) is improper.

With only an intercept term ($\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_1$), Chapter 5 of Gelman et al. (2013) specifies that the joint posterior density function $\pi_{\text{hyp.post}}(r, \beta_1 \mid \mathbf{y})$ with $dr/r^{1.5}$ and independently with the proper standard logistic distribution on β_1 is proper if there is at least one interior group. However, the resulting posterior can be improper with this condition. For example, when there is only one interior group with two trials ($y_1 = 1, n_1 = 2$), the joint posterior density function of r and β_1 is

$$\pi_{\text{hyp.post}}(r, \beta_1 \mid \mathbf{y}) \propto \pi_{\text{hyp.prior}}(r, \beta_1) \times L(r, \beta_1) \propto \frac{p^E q^E}{r^{1.5}} \times \frac{r p^E q^E}{(1+r)}, \quad (1.19)$$

where $p^E = \exp(\beta_1)/(1 + \exp(\beta_1)) = 1 - q^E$. The integration of this joint posterior density function with respect to r is not finite because the density function goes to infinity as r approaches zero. (The integral of $dr/r^{0.5}$ over the range $[0, 0 + \epsilon]$ for a positive constant ϵ is not finite.) To achieve posterior propriety in this setting, we need at least two interior groups in the data as shown in Figure 1.1.

The posterior distributions of Kass and Steffey (1989) and Kahn and Raftery (1996) are always improper regardless of the data due to their hyper-prior PDF dr/r . This is because the likelihood function in (1.7) approaches $c(\boldsymbol{\beta})$, a positive constant with respect to r , as r increases to infinity. Then the hyper-prior PDF dr/r , whose integration becomes infinite over the range $[\epsilon, \infty)$ for a positive constant ϵ , governs the right tail behavior of the conditional posterior density function of r , $\pi_{\text{hyp.cond.post}}(r \mid \boldsymbol{\beta}, \mathbf{y})$. It indicates that $\pi_{\text{hyp.cond.post}}(r \mid \boldsymbol{\beta}, \mathbf{y})$ is improper, and thus the joint posterior density $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is improper.

1.3.3 Inference when a posterior distribution is improper

Making an inference based on an improper posterior distribution is dangerous because most statistical inferential tools assume that the target distribution is a probability distribution but the improper posterior distribution is not a probability distribution. For example, Hobert

and Casella (1996) call attention to running a Gibbs sampler on an improper posterior distribution because the Gibbs sampler may seem to work well even when the posterior distribution is improper. They emphasize checking posterior propriety in advance to prevent a (non-recurrent) Gibbs chain from converging to some nonexistent probability distribution. Athreya and Roy (2014) also show that Markov chain Monte Carlo methods can be misleading when the posterior is improper because a standard average estimator based on Markov chains converges to zero with probability one. They introduce regenerative sequence Monte Carlo methods that enable a valid inference even when a posterior distribution is improper.

When it comes to a BBL model, the conditions for posterior propriety in Figure 1.1 can be met in most cases because in practice the data are composed of a suitably large number of groups, k . However, improper hyper-prior PDFs may result in posterior impropriety when the data are composed of a small number of groups. In this case, we recommend using proper hyper-prior PDFs for r and $\boldsymbol{\beta}$, e.g., a uniform shrinkage prior on r , $dr/(t+r)^2$, which is known to produce good frequentist properties (Strawderman, 1971; Morris and Christiansen, 1997), and a diffuse Gaussian prior on $\boldsymbol{\beta}$ with relatively large standard deviations (Kahn and Raftery, 1996). Setting a small constant t in a uniform shrinkage prior is considered as a conservative choice that allows the data to speak more with smaller shrinkage factors (Morris and Christiansen, 1997). Another possibility (except when $n_j = 1$ for all j) is to estimate MLEs of r and $\boldsymbol{\beta}$ via (1.7) and plug these estimates into the conditional Beta distributions of random effects in (1.4). This approach can be considered as an empirical Bayes (EB) approach (Efron and Morris, 1975) with $\pi_{\text{hyp.prior}}(r, \boldsymbol{\beta}) \propto 1$. However, this EB approach tends to be over-confident in estimating random effects when k is small because the EB approach does not account for the uncertainties of unknown r and $\boldsymbol{\beta}$ though these uncertainties are large when k is small.

1.3.4 Numerical illustration: Data of two bent coins

We have two biased coins; a bent penny and a possibly differently bent nickel ($k = 2$). We flip these coins twice for each ($n_1 = n_2 = 2$) and record the number of Heads for the penny (y_1) and also for the nickel (y_2). We model this experiment as $y_j | p_j \sim \text{Bin}(2, p_j)$ independently, where p_j is the unknown probability of observing Heads for coin j . We assume an i.i.d. prior distribution for random effects, $p_j | r, \beta_1 \sim \text{Beta}(rp^E, rq^E)$, where $p^E = \exp(\beta_1)/[1 + \exp(\beta_1)] = 1 - q^E$, i.e., a BB model.

We look into posterior propriety under four different settings depending on whether the hyper-prior distribution for β_1 (or equivalently p^E) is proper or improper flat $d\beta$, and on whether the hyper-prior distribution of r is proper or dr/r^2 .

Table 1.1 shows when the posterior distribution is proper (denoted by O) and when it is not (denoted by X). The posterior distribution in case (a) is always proper because both hyper-prior distributions for r and β_1 are proper. In case (b) where β_1 has the Lebesgue measure and r has a proper hyper-prior PDF, the posterior is proper unless both coins land either all Heads ($y_1 = y_2 = 2$) or all Tails ($y_1 = y_2 = 0$). This is because the condition for posterior propriety is that the covariate matrix of interior coins is of full rank and this condition without any covariates is met if at least one coin is interior; see Figure 1.1. In

(a) Any proper $f(r)$ and any proper $g(\beta_1)$

| $y_1 \backslash y_2$ | 0 | 1 | 2 |
|----------------------|---|---|---|
| 0 | O | O | O |
| 1 | O | O | O |
| 2 | O | O | O |

(c) $f(r) \propto 1/r^2$ and any proper $g(\beta_1)$

| $y_1 \backslash y_2$ | 0 | 1 | 2 |
|----------------------|---|---|---|
| 0 | X | X | X |
| 1 | X | O | X |
| 2 | X | X | X |

(b) Any proper $f(r)$ and $g(\beta_1) \propto 1$

| $y_1 \backslash y_2$ | 0 | 1 | 2 |
|----------------------|---|---|---|
| 0 | X | O | O |
| 1 | O | O | O |
| 2 | O | O | X |

(d) $f(r) \propto 1/r^2$ and $g(\beta_1) \propto 1$

| $y_1 \backslash y_2$ | 0 | 1 | 2 |
|----------------------|---|---|---|
| 0 | X | X | X |
| 1 | X | O | X |
| 2 | X | X | X |

Table 1.1: The symbol O indicates that the posterior distribution is proper on corresponding data, and the symbol X indicates that the posterior distribution is not proper on corresponding data.

cases (c) and (d), where r has the improper hyper-prior PDF, dr/r^2 , posterior propriety holds only when each coin shows one Head and one Tail, i.e., both coins are interior ($y_1 = y_2 = 1$); see Figure 1.1. Cases (c) and (d) have the same condition for posterior propriety because the condition that arises from the improper flat hyper-prior PDF for β_1 in case (d) is automatically met if the condition arising from the improper hyper-prior PDF for r , i.e., $k_y \geq 2$, is met.

Next, we check the effect of different joint hyper-prior PDFs used in cases (a)–(d) on the random effect estimation, e.g., p_1 . For this purpose, we set $g(\beta_1) = N(\beta_1 | 0, 10^{10})$, a diffuse Gaussian distribution with mean zero and variance 10^{10} for a proper hyper-prior PDF of β_1 , and set $f(r) \propto 1/(10^{-5} + r)^2$ for a proper hyper-prior PDF of r . We draw 55,000 posterior samples of r and β_1 from their joint posterior distribution, $\pi_{\text{hyp.post}}(r, \beta_1 | \mathbf{y})$, using a Metropolis within Gibbs sampler (Tierney, 1994a), discarding the first 5,000 samples as burn-ins. We adjust proposal scales of independent Gaussian proposal distributions to obtain a reasonable acceptance probability around 0.35 for each parameter. Using the posterior samples of r and β_1 , we draw the posterior sample of p_1 from its marginal posterior distribution $\pi_{\text{marg.post}}(p_1 | \mathbf{y})$ via a Monte Carlo integration:

$$\pi_{\text{marg.post}}(p_1 | \mathbf{y}) = \int_{\mathbf{R}} \int_0^\infty \pi_{\text{cond.post}}(p_1 | r, \beta_1, \mathbf{y}) \times \pi_{\text{hyp.post}}(r, \beta_1 | \mathbf{y}) dr d\beta_1, \quad (1.20)$$

i.e., sampling p_1 from $\pi_{\text{cond.post}}(p_1 | r, \beta_1, \mathbf{y})$ given already sampled r and β_1 . In addition, we estimate p_1 via an EB approach for a comparison; estimating MLEs of r and β_1 , inserting these into $\pi_{\text{cond.post}}(p_1 | r, \beta_1, \mathbf{y})$, and calculating (0.025, 0.975) quantiles of this conditional Beta posterior distribution $\pi_{\text{cond.post}}(p_1 | r, \beta_1, \mathbf{y})$.

We fit these models on the data $\{y_1 = y_2 = 1\}$ for which posterior distributions in cases (a)–(d) are all proper. The resulting 95% posterior intervals for p_1 are summarized in the first row of Table 1.2. All these intervals are similar because the proper hyper-prior PDF of r , $dr/(10^{-5} + r)^2$, used in cases (a) and (b) mimics well its improper limit, dr/r^2 , used in cases (c) and (d), and because the diffuse Gaussian hyper-prior PDF of β_1 behaves similarly to an improper flat density function. These intervals are wide, reflecting on the lack of

| Data \ Model | Case (a) | Case (b) | Case (c) | Case (d) | EB |
|--------------------|----------------|----------------|----------------|----------------|----------------|
| $y_1 = y_2 = 1$ | (0.048, 0.950) | (0.048, 0.951) | (0.049, 0.951) | (0.049, 0.950) | (0.490, 0.510) |
| $y_1 = 0, y_2 = 1$ | (0.000, 0.247) | (0.000, 0.242) | Improper | Improper | (0.218, 0.284) |

Table 1.2: The 95% posterior intervals of p_1 obtained by Bayesian BBL models equipped with joint hyper-prior PDFs in cases (a)–(d), and those obtained by an empirical Bayes (EB) approach. We set $g(\beta_1) = N(\beta_1 | 0, 10^{10})$ and $f(r) \propto 1/(10^{-5} + r)^2$ for proper hyper-prior PDFs of β_1 and r , respectively.

information about r and β_1 in two observations. However, the EB interval centered at 0.5 is much too narrow because it does not account for the uncertainties of unknown r and β_1 .

The hyper-prior PDFs in cases (c) and (d) result in an improper posterior for the data $\{y_1 = 0, y_2 = 1\}$. Thus, we fit models equipped with hyper-prior PDFs in cases (a) and (b) and an EB model on these data. The posterior intervals for p_1 are summarized in the second row of Table 1.2. The intervals in cases (a) and (b) are similar because the diffuse Gaussian prior for β_1 is close to an improper flat prior. The EB interval centered at around 0.25 is again much narrower than the full Bayesian intervals in (a) and (b).

1.3.5 Numerical illustration: Data of five hospitals

New York State Cardiac Advisory Committee (2014) has reported the outcomes for the Valve Only and Valve/CABG surgeries. The data are based on the patients discharged between December 1, 2008, and November 30, 2011 in 40 non-federal hospitals in New York State. We select the smallest five hospitals with respect to the number of patients for simplicity. Table 1.3 shows the data including the number of cases (n_j), the number

| j | 1 | 2 | 3 | 4 | 5 |
|----------|------|------|------|------|------|
| n_j | 54 | 75 | 93 | 104 | 105 |
| y_j | 3 | 4 | 1 | 1 | 1 |
| EMR $_j$ | 4.30 | 2.21 | 2.59 | 4.73 | 3.28 |

Table 1.3: Data of five hospitals. The number of patients in hospital j is denoted by n_j , the number of death in hospital j is denoted by y_j , and the expected mortality rate (%) for hospital j is denoted by EMR $_j$.

of deaths (y_j), and expected mortality rate (EMR_j). The EMR_j is a hospital-wise average over the predicted probabilities of death for each patient; the larger the EMR_j is, the more difficult cases hospital j handles. We use the EMR_j as a continuous covariate. We assume $y_j | p_j \sim \text{Bin}(n_j, p_j)$ independently. We also assume that the unknown true mortality rates p_j come from independent conjugate Beta prior distributions in (1.2) with $x_j^T \boldsymbol{\beta} = \beta_1 x_{1j} + \beta_2 x_{2j}$, where $x_{1j} = 1$ and $x_{2j} = \text{EMR}_j$.

We consider four joint hyper-prior densities: $d\boldsymbol{\beta}dr/r^2$, $d\boldsymbol{\beta}dr/(1+r)^2$, $d\boldsymbol{\beta}dr/r^{1.5}$ and $d\boldsymbol{\beta}dr/(1+r)^{1.5}$. The condition for posterior propriety is $\text{rank}(X_y) = 2$ for all four joint hyper-prior PDFs because this condition automatically meets $k_y \geq 2$. The data in Table 1.3 satisfy the condition for posterior propriety because all the hospitals are interior ($1 \leq y_j \leq n_j - 1$ for all j and thus $k = k_y = 5$) and their covariate matrix $X = X_y$ is of full rank.

Based on the data in Table 1.3, we make two hypothetical data sets in Table 1.4. In the first hypothetical data set, only one hospital is interior. The resulting posterior distribution is improper for the four joint hyper-prior PDFs because the rank of the covariate matrix of this interior hospital is not two ($\text{rank}(X_y) = 1$). In the second hypothetical data set, two hospitals are interior but their EMRs are the same, meaning that the rank of the covariate matrix of these two interior hospitals is one. Thus, the resulting posterior is improper for the four joint hyper-prior PDFs.

Next we compare several models using these data sets in Table 1.3 and Table 1.4 to see the effect of different constants, t and u , in $dr/(t+r)^{u+1}$; we consider using either $u = 1$ or

| | | | | | | | | | | | |
|----------------|------|------|------|------|------|----------------|------|------|------|------|------|
| j | 1 | 2 | 3 | 4 | 5 | j | 1 | 2 | 3 | 4 | 5 |
| n_j | 54 | 75 | 93 | 104 | 105 | n_j | 54 | 75 | 93 | 104 | 105 |
| y_j | 1 | 0 | 0 | 0 | 0 | y_j | 1 | 2 | 0 | 0 | 0 |
| EMR_j | 4.30 | 2.21 | 2.59 | 4.73 | 3.28 | EMR_j | 4.30 | 4.30 | 2.59 | 4.73 | 3.28 |

Table 1.4: Two hypothetical data sets of five hospitals. The number of patients in hospital j is denoted by n_j , the number of death in hospital j is denoted by y_j , and the expected mortality rate (%) for hospital j is denoted by EMR_j . In the first data set, only the first hospital is interior. In the second data set, the first two hospitals are interior but their EMRs are the same.

| Data\Model | $1/r^2$ | $1/(10^{-5} + r)^2$ | $1/r^{1.5}$ | $1/(10^{-5} + r)^{1.5}$ | EB |
|---------------|----------------|---------------------|----------------|-------------------------|----------------|
| Table 1.3 | (0.011, 0.116) | (0.011, 0.115) | (0.008, 0.099) | (0.008, 0.100) | (0.012, 0.046) |
| Table 1.4 (L) | Improper | (0.000, 0.067) | Improper | (0.000, 0.066) | (0.003, 0.005) |
| Table 1.4 (R) | Improper | (0.000, 0.068) | Improper | (0.001, 0.062) | (0.002, 0.030) |

Table 1.5: The 95% posterior intervals of p_1 obtained by Bayesian BBL models equipped with hyper-prior PDFs, $g(\boldsymbol{\beta}) = N(\boldsymbol{\beta} \mid 0 \times \mathbf{1}_2, 10^{10} \times I_2)$, which is the same for all models, and $dr/(t+r)^{u+1}$ with $u = 1$ or $u = 0.5$ and with $t = 0$ or $t = 10^{-5}$. The 95% intervals obtained by an empirical Bayes (EB) approach appear in the last column. The left hypothetical data in Table 1.4 are denoted by Table 1.4 (L) and the right one by Table 1.4 (R).

$u = 0.5$ and either $t = 0$ or $t = 10^{-5}$. The sampling configurations are the same as those in the previous section except that we set $g(\boldsymbol{\beta}) = N(\boldsymbol{\beta} \mid 0 \times \mathbf{1}_2, 10^{10} \times I_2)$ for all models, where $\mathbf{1}_2$ is a vector of ones and I_2 is a 2×2 identity matrix. Table 1.5 summarizes the 95% posterior intervals for p_1 .

When models are all proper based on the data in Table 1.3, the interval estimates are similar between $t = 10^{-5}$ and $t = 0$, but quite different depending on whether $u = 1$ or $u = 0.5$. Clearly, intervals based on $u = 1$ are wider (more conservative) than those based on $u = 0.5$. This is because dr/r^2 puts more weight at zero than $dr/r^{1.5}$ a priori, and thus dr/r^2 produces smaller posterior samples of r that leads to wider interval estimates in turn; the variance of a conditional Beta posterior distribution for p_j in (1.4), $\hat{p}_j(1 - \hat{p}_j)/(r + n_j + 1)$, increases as r decreases, where \hat{p}_j is its posterior mean. The improper hyper-prior PDFs, dr/r^2 and $dr/r^{1.5}$, lead to posterior impropriety for the data in Table 1.4 due to the reasons specified above. The EB approach leads to much narrower intervals for all three data sets.

1.4 Acceptance-rejection method

In this section, we illustrate an acceptance-rejection (A-R) method to draw posterior samples of random effects and hyper-parameters (Robert and Casella, 2013). The joint posterior density function of $\alpha = -\log(r)$ and $\boldsymbol{\beta}$ based on their joint hyper-prior density function in (1.9) is

$$f(\alpha, \boldsymbol{\beta} \mid \mathbf{y}) \propto f(\alpha, \boldsymbol{\beta})L(\alpha, \boldsymbol{\beta}) \propto \exp(\alpha)L(\alpha, \boldsymbol{\beta}). \quad (1.21)$$

The A-R method is useful when it is difficult to sample a parameter of interest θ directly from its target probability density $f(\theta)$, which is known up to a normalizing constant, but an easy-to-sample envelope function $g(\theta)$ is available. The A-R method samples θ from the envelope $g(\theta)$ and accepts it with a probability $f(\theta)/(Mg(\theta))$, where M is a constant making $f(\theta)/g(\theta) \leq M$ for all θ . The distribution of the accepted θ exactly follows $f(\theta)$. The A-R method is stable as long as the tails of the envelope function are thicker than those of the target density function.

Thus, our goal is to draw posterior samples of hyper-parameters from (1.21), using an easy-to-sample envelope function $g(\alpha, \boldsymbol{\beta})$ that has thicker tails than the target density function. We factor the envelope function into two parts, $g(\alpha, \boldsymbol{\beta}) = g_1(\alpha)g_2(\boldsymbol{\beta})$ to model the tails of each function separately. We consider the tail behavior of the conditional posterior density function $f(\alpha | \boldsymbol{\beta}, \mathbf{y})$ to establish $g_1(\alpha)$; $f(\alpha | \boldsymbol{\beta}, \mathbf{y})$ behaves as $\exp(-\alpha(k-1))$ when α goes to ∞ and as $\exp(\alpha)$ when α goes to $-\infty$. It indicates that $f(\alpha | \boldsymbol{\beta}, \mathbf{y})$ is skewed to the left because the right tail touches the x -axis faster than the left tail does as long as $k > 1$. A skewed t -distribution is a good candidate for $g_1(\alpha)$ because it behaves as a power law on both tails, leading to thicker tails than those of $f(\alpha | \boldsymbol{\beta}, \mathbf{y})$.

It is too complicated to figure out the tail behaviors of $f(\boldsymbol{\beta} | \alpha, \mathbf{y})$. However, because $f(\boldsymbol{\beta} | \alpha, \mathbf{y})$ in the Gaussian model (as an approximation) has a multivariate Gaussian density function (Morris and Tang, 2011a; Kelly, 2014), we consider a multivariate t -distribution with four degrees of freedom as a good candidate for $g_2(\boldsymbol{\beta})$.

Specifically, we assume

$$g_1(\alpha) = g_1(\alpha; l, \sigma, a, b) \equiv \text{Skewed-}t(\alpha | l, \sigma, a, b), \quad (1.22)$$

$$g_2(\boldsymbol{\beta}) = g_2(\boldsymbol{\beta}; \boldsymbol{\xi}, S_{(m \times m)}) \equiv t_4(\boldsymbol{\beta} | \boldsymbol{\xi}, S), \quad (1.23)$$

where $\text{Skewed-}t(\alpha | l, \sigma, a, b)$ represents a density function of a skewed t -distribution of α with location l , scale σ , degree of freedom $a + b$, and skewness $a - b$ for any positive constants a

and b (Jones and Faddy, 2003). Jones and Faddy (2003) derive the mode of $g_1(\alpha)$ as

$$l + \frac{(a-b)\sqrt{a+b}}{\sqrt{(2a+1)(2b+1)}}, \quad (1.24)$$

and provide a representation to generate random variables that follows Skewed- $t(\alpha \mid l, \sigma, a, b)$;

$$\alpha \sim l + \sigma \frac{\sqrt{a+b}(2T-1)}{2\sqrt{T(1-T)}}, \text{ where } T \sim \text{Beta}(a, b). \quad (1.25)$$

They also show that the tails of the skewed- t density function follow a power law with $\alpha^{-(2a+1)}$ on the left and $\alpha^{-(2b+1)}$ on the right when $b > a$.

The notation $t_4(\boldsymbol{\beta} \mid \boldsymbol{\xi}, S)$ in (1.23) indicates a density function of a multivariate t -distribution of $\boldsymbol{\beta}$ with four degrees of freedom, a location vector $\boldsymbol{\xi}$, and a $m \times m$ scale matrix S that leads to the variance-covariance matrix $2S$.

We determine the parameters of $g_1(\alpha)$ and $g_2(\boldsymbol{\beta})$, *i.e.*, l , σ , a , b , $\boldsymbol{\xi}$, and S , to make the product of $g_1(\alpha)$ and $g_2(\boldsymbol{\beta})$ similar to the target joint posterior density $f(\alpha, \boldsymbol{\beta} \mid \mathbf{y})$. First, we obtain the mode of $f(\alpha, \boldsymbol{\beta} \mid \mathbf{y})$, $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$, and the inverse of the negative Hessian matrix at the mode. We define $-H_{\hat{\alpha}}^{-1}$ to indicate the (1, 1)th element of the negative Hessian matrix and $-H_{\hat{\boldsymbol{\beta}}}^{-1}$ to represent the negative Hessian matrix without the first row and the first column.

For $g_1(\alpha)$, we set (a, b) to $(k, 2k)$ if k is less than 10 (or to $(\log(k), 2\log(k))$ otherwise) for a left-skewness and these small values of a and b lead to thick tails. We match the mode of $g_1(\alpha)$ specified in (1.24) to $\hat{\alpha}$ by setting the location parameter l to $\hat{\alpha} - (a - b)\sqrt{a+b}/\sqrt{(2a+1)(2b+1)}$. We set the scale parameter σ to $(-H_{\hat{\alpha}}^{-1})^{0.5}\psi$, where ψ is a tuning parameter; when the A-R method produces extreme weights defined in (1.26) below, we need enlarge the value of ψ .

For $g_2(\boldsymbol{\beta})$, we set the location vector $\boldsymbol{\xi}$ to the mode $\hat{\boldsymbol{\beta}}$ and the scale matrix S to $-H_{\hat{\boldsymbol{\beta}}}^{-1}/2$ so that the variance-covariance matrix becomes $-H_{\hat{\boldsymbol{\beta}}}^{-1}$.

For implementation of the acceptance-rejection method, we draw four times more trial samples than the desired number of samples, denoted by N , independently from $g_1(\alpha)$ and $g_2(\boldsymbol{\beta})$. We calculate $4N$ weights, each of which is defined as

$$w_i \equiv w(\alpha^{(i)}, \boldsymbol{\beta}^{(i)}) = \frac{f(\alpha^{(i)}, \boldsymbol{\beta}^{(i)} \mid \mathbf{y})}{g_1(\alpha^{(i)})g_2(\boldsymbol{\beta}^{(i)})}, \text{ for } i = 1, 2, \dots, 4N. \quad (1.26)$$

The A-R method accepts each pair of $(\alpha^{(i)}, \boldsymbol{\beta}^{(i)})$ with a probability w_i/M where M is set to the maximum of all the $4N$ weights. When the A-R method accepts more than N pairs, it discards the redundant. If the A-R method accepts less than N pairs, then it additionally draws N' (six times the shortage) pairs and calculates a new maximum M' from all the previous and new weights; the A-R method accepts or rejects the entire pairs again with new probabilities w_j/M' , $j = 1, 2, \dots, 4N + N'$.

After obtaining posterior samples of hyper-parameters, we draw posterior samples of random effects from

$$f(\mathbf{p} \mid \mathbf{y}) = \int f(\mathbf{p} \mid r, \boldsymbol{\beta}, \mathbf{y}) f(r, \boldsymbol{\beta} \mid \mathbf{y}) dr d\boldsymbol{\beta}, \quad (1.27)$$

where $\mathbf{p} = (p_1, p_2, \dots, p_k)^\top$ and the distributions in the integrand are given in (1.4) and (1.9). The integration on the right hand side of (1.27) can be done by sampling \mathbf{p} from $f(p_j \mid \boldsymbol{\beta}, r, \mathbf{y})$ in (1.4) for $j = 1, 2, \dots, k$, given $r = \exp(-\alpha)$ and $\boldsymbol{\beta}$ that are already sampled from $f(\alpha, \boldsymbol{\beta} \mid \mathbf{y})$, or equivalently $f(r, \boldsymbol{\beta} \mid \mathbf{y})$, via the A-R method.

1.5 Frequency method checking

The question as to whether the interval estimates of random effects for given confidence level obtained by a specific model achieve the nominal coverage rate for any true parameter values is one of the key model evaluation criteria. Unlike standard model checking methods that test whether a two-level model is appropriate for data (Dean, 1992; Christiansen and Morris, 1996), frequency method checking is a procedure to evaluate the coverage properties of the model. Conditioning that the two-level model is appropriate, the frequency method checking generates pseudo-data sets given specific values of hyper-parameters and estimates unknown coverage probabilities based on these mock data sets (a parametric bootstrapping).

1.5.1 Pseudo-data generation

Figure 1.2 displays the process of generating pseudo-data sets. It is noted that the conjugate prior distribution of each random effect in (1.2) is completely determined by two hyperparameters, r and β . Fixing r and β at specific values, we generate N_{sim} sets of random effects from the conjugate prior distribution, *i.e.*, $\{\mathbf{p}^{(i)}, i = 1, \dots, N_{\text{sim}}\}$, where the superscript (i) indicates the i -th simulation. Next, using the distribution of observed data in (1.1), we generate N_{sim} sets of observed data sets $\{\mathbf{y}^{(i)}, i = 1, \dots, N_{\text{sim}}\}$ given each $\mathbf{p}^{(i)}$.

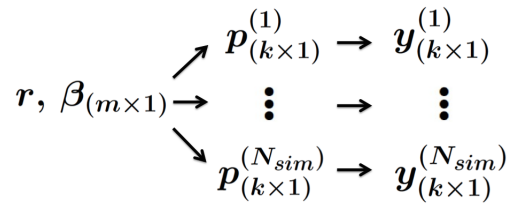


Figure 1.2: Pseudo-data generating process.

1.5.2 Coverage probability estimation

After fitting the Gaussian model for each simulated data set, we obtain interval estimates of the random effects $\mathbf{p}^{(i)}$. Let $(\hat{p}_{j, \text{low}}^{(i)}, \hat{p}_{j, \text{upp}}^{(i)})$ represent the lower and upper bounds of the interval estimate of random effect j based on the i -th simulation given a specific confidence level. We define the coverage indicator of random effect j on the i -th mock data set as

$$I(p_j^{(i)}) = \begin{cases} 1, & \text{if } p_j^{(i)} \in (\hat{p}_{j, \text{low}}^{(i)}, \hat{p}_{j, \text{upp}}^{(i)}), \\ 0, & \text{otherwise.} \end{cases} \quad (1.28)$$

This shrinkage indicator is equal to the value one if the random effect j in simulation i is between its interval estimates and zero otherwise.

1.5.3 Simple unbiased coverage estimator.

When the confidence level is 95%, the proportion of 95% interval estimates that contain random effect j is an intuitive choice for the coverage rate estimator for random effect

j . This estimator implicitly assumes that there exist k unknown coverage probabilities of random effects, denoted by $C_{r,\boldsymbol{\beta}}(p_j)$ for $j = 1, 2, \dots, k$, depending on the values of the hyperparameters that generate random effects and mock data sets. The coverage indicators for random effect j in (1.28) is assumed to follow an independent and identically distributed Bernoulli distribution given the unknown coverage rate $C_{r,\boldsymbol{\beta}}(p_j)$. The sample mean of these coverage indicators is a simple unbiased coverage estimator for $C_{r,\boldsymbol{\beta}}(p_j)$; for $j = 1, 2, \dots, k$,

$$\bar{I}(p_j) = \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} I(p_j^{(i)}). \quad (1.29)$$

The unbiased variance estimator of $\text{Var}(\bar{I}(p_j))$ is, for $j = 1, 2, \dots, k$,

$$\widehat{\text{Var}}(\bar{I}(p_j)) = \frac{1}{N_{\text{sim}}(N_{\text{sim}} - 1)} \sum_{i=1}^{N_{\text{sim}}} \left(I(p_j^{(i)}) - \bar{I}(p_j) \right)^2. \quad (1.30)$$

1.5.4 Rao-Blackwellized unbiased coverage estimator.

Frequency method checking is computationally expensive in nature because it fits a model on every mock data set. The situation deteriorates if the number of simulations or the size of data is large, or the estimation method is computationally demanding. Christiansen and Morris (1997) and Tang (2002) use a Rao-Blackwellized (RB) unbiased coverage estimator for the unknown coverage rate of each random effect, which is more efficient than the simple unbiased coverage estimator. For $j = 1, 2, \dots, k$,

$$C_{r,\boldsymbol{\beta}}(p_j) = E(\bar{I}(p_j) \mid r, \boldsymbol{\beta}) = E \left[\frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} E \left(I(p_j^{(i)}) \mid r, \boldsymbol{\beta}, \mathbf{y}^{(i)} \right) \mid r, \boldsymbol{\beta} \right], \quad (1.31)$$

where the sample mean of the interior conditional expectations in (1.31) is the RB unbiased coverage estimator. Specifically,

$$\bar{I}^{RB}(p_j) = \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} E \left(I(p_j^{(i)}) \mid r, \boldsymbol{\beta}, \mathbf{y}^{(i)} \right) \quad (1.32)$$

$$= \frac{1}{N_{\text{sim}}} \sum_{i=1}^{N_{\text{sim}}} P \left(p_j^{(i)} \in (\hat{p}_{j, \text{low}}^{(i)}, \hat{p}_{j, \text{upp}}^{(i)}) \mid r, \boldsymbol{\beta}, \mathbf{y}^{(i)} \right). \quad (1.33)$$

We can easily compute the conditional posterior probabilities in (1.33) using the cumulative density function of the Gaussian conditional posterior distribution of each random effect in (1.4). The variance of $\bar{I}^{RB}(p_j)$ does not exceed the variance of a simple unbiased coverage estimator, $\bar{I}(p_j)$ (Rao, 1945; Blackwell, 1947).

If one dataset $\mathbf{y}^{(i)}$ is simulated for each set of random effects $\mathbf{p}^{(i)}$, the variance estimator below is an unbiased estimator of $Var(\bar{I}^{RB}(p_j))$. For $j = 1, 2, \dots, k$,

$$\widehat{Var}(\bar{I}^{RB}(p_j)) \equiv \frac{1}{N_{\text{sim}}(N_{\text{sim}} - 1)} \sum_{i=1}^{N_{\text{sim}}} \left(E(I(p_j^{(i)})) \mid r, \boldsymbol{\beta}, \mathbf{y}^{(i)} - \bar{I}^{RB}(p_j) \right)^2. \quad (1.34)$$

1.5.5 Overall unbiased coverage estimator

To summarize the frequency method checking, we report the overall unbiased coverage estimate and its variance estimate,

$$\bar{\bar{I}}^{RB} = \frac{1}{k} \sum_{j=1}^k \bar{I}^{RB}(p_j) \quad \text{and} \quad \widehat{Var}(\bar{\bar{I}}^{RB}) = \frac{1}{k^2} \sum_{j=1}^k \widehat{Var}(\bar{I}^{RB}(p_j)). \quad (1.35)$$

1.6 Example: Data of 18 baseball players

Table 1.6 shows the data of 18 major league baseball players through their first 45 official at-bats in the 1970 season (Efron and Morris, 1975). Our goal is to obtain point and interval estimates of each player's unknown true batting average p_j (random effect) for a comparison purpose, whilst considering a binary covariate information about whether a player was an outfielder or not. We assume that each player's unknown true batting average did not change for the first 45 at-bats, and the at-bats were independent to each other given the unknown true batting average.

We fit a Bayesian BBLR model ($m = 2$) to these data with the SHP analog, $f_1(r, \boldsymbol{\beta}) \propto d\boldsymbol{\beta}dr/r^2$, for an illustrative purpose and three other joint HPDFs, $f_2(r, \boldsymbol{\beta}) \propto d\boldsymbol{\beta}dr/(1+r)^2$, $f_3(r, \boldsymbol{\beta}) \propto d\boldsymbol{\beta}dr/r^{1.5}$, and $f_4(r, \boldsymbol{\beta}) \propto d\boldsymbol{\beta}dr/(1+r)^{1.5}$, for a sensitivity analysis.

We summarize 5,000 posterior samples obtained via the A-R method in Table 1.6. The binary covariate information forms two different conjugate Beta prior distributions for out-

Table 1.6: Data of 18 baseball players based on the first 45 official at-bats in the 1970 season and summaries of 5,000 posterior samples obtained via the A-R method with $d\text{rd}\boldsymbol{\beta}/r^2$; for player j , \bar{y}_j is the observed batting average out of 45 at-bats, x_{j2} is an outfielder indicator taking on 1 if player j is an outfielder and 0 otherwise, $E(B_j|\mathbf{y})$ is the posterior mean of shrinkage factor, $E(p_j^E|\mathbf{y})$ is the posterior mean of $\text{logit}^{-1}(\mathbf{x}_j^\top \boldsymbol{\beta})$, $E(p_j|\mathbf{y})$ is the posterior mean of random effect j , and 95% P.I. is the (0.025, 0.975) quantiles of 5,000 posterior samples of random effect j .

| Names | \bar{y}_j | x_{j2} | $E(B_j \mathbf{y})$ | $E(p_j^E \mathbf{y})$ | $E(p_j \mathbf{y})$ | 95% P.I. |
|------------------|-------------|----------|---------------------|-----------------------|---------------------|----------------|
| Roberto Clemente | 0.400 | 1 | 0.752 | 0.309 | 0.332 | (0.257, 0.429) |
| Frank Robinson | 0.378 | 1 | 0.752 | 0.309 | 0.324 | (0.249, 0.416) |
| Frank Howard | 0.356 | 1 | 0.752 | 0.309 | 0.321 | (0.244, 0.411) |
| Jay Johnstone | 0.333 | 1 | 0.752 | 0.309 | 0.315 | (0.236, 0.401) |
| Ken Berry | 0.311 | 1 | 0.752 | 0.309 | 0.309 | (0.233, 0.394) |
| Ron Swaboda | 0.244 | 1 | 0.752 | 0.309 | 0.293 | (0.209, 0.372) |
| Del Unser | 0.222 | 1 | 0.752 | 0.309 | 0.288 | (0.201, 0.370) |
| Billy Williams | 0.222 | 1 | 0.752 | 0.309 | 0.286 | (0.197, 0.367) |
| Jim Spencer | 0.311 | 0 | 0.752 | 0.233 | 0.252 | (0.186, 0.341) |
| Don Kessinger | 0.289 | 0 | 0.752 | 0.233 | 0.246 | (0.179, 0.331) |
| Luis Alvarado | 0.267 | 0 | 0.752 | 0.233 | 0.241 | (0.175, 0.321) |
| Ron Santo | 0.244 | 0 | 0.752 | 0.233 | 0.235 | (0.167, 0.311) |
| Rico Petrocelli | 0.222 | 0 | 0.752 | 0.233 | 0.229 | (0.162, 0.301) |
| Ellie Rodriguez | 0.222 | 0 | 0.752 | 0.233 | 0.230 | (0.163, 0.305) |
| George Scott | 0.222 | 0 | 0.752 | 0.233 | 0.229 | (0.161, 0.304) |
| Bert Campaneris | 0.200 | 0 | 0.752 | 0.233 | 0.225 | (0.155, 0.296) |
| Thurman Munson | 0.178 | 0 | 0.752 | 0.233 | 0.219 | (0.146, 0.288) |
| Max Alvis | 0.156 | 0 | 0.752 | 0.233 | 0.213 | (0.139, 0.281) |

fielders and infielders. The posterior mean of β_1 is 0.386 with its (0.025, 0.975) quantiles equal to (0.004, 0.760). As a result, the posterior mean of outfielder's expected batting average, i.e., $E(p_j^E|\mathbf{y})$, is larger than that for infielder's (0.309 > 0.233).

We draw the 95% posterior interval plot in Figure 1.3. The result clearly shows an effect called *regression towards the mean* (RTTM) within outfielders and infielders. The first player, for example, is an outfielder and his observed batting average is higher than any other outfielders. This can be attributed to his good luck, considering that his observed batting average is close to the upper bound of the posterior interval estimate. The RTTM implies that his unknown true batting average will shrink towards outfielder's expected batting average in the long run. Thus, reflecting on the RTTM, the posterior mean of his unknown true batting average becomes lower than his observed batting average.

We check the sensitivity of the posterior inference according to different hyper-prior

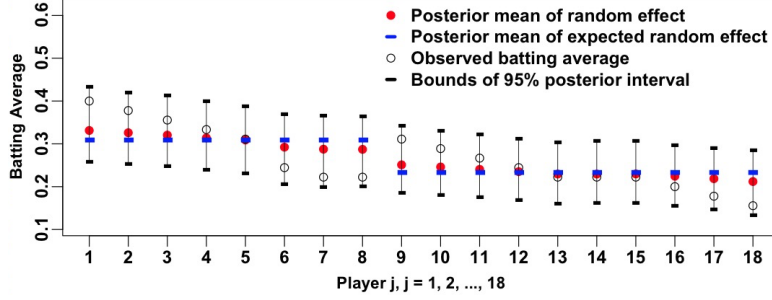


Figure 1.3: The posterior means (red dots) of unknown true batting averages (random effects) shrink the observed batting averages (empty circles) towards the posterior mean of expected batting averages $E(p_j^E | \mathbf{y})$ (blue horizontal bars). The 95% posterior intervals (vertical lines) clarify the regression towards the mean.

distributions and summarize the outcomes in Table 1.7. The posterior inference on logistic regression coefficients seems robust to the different HPDFs of r . The inference on r that determines shrinkage factors is quite sensitive to the exponent (2 or 1.5) of the HPDF for r ; the posterior mean of the shrinkage factor is 0.752 when dr/r^2 is used and 0.862 when $dr/r^{1.5}$ is used. The inference is more conservative when the exponent is two in that the posterior distribution borrows less information (smaller shrinkage) from the second-level prior distribution. The inference result looks insensitive to whether $f(r)$ is improper ($t = 1$) or not ($t = 0$), considering that the posterior mean of the shrinkage factor is 0.752 for dr/r^2

Table 1.7: The posterior means (P.M.), posterior standard deviations (P.SD.), and 95% posterior intervals (95% P.I.) of $B = r/(r + 45)$, β_1 , and β_2 according to different hyper-prior distributions.

| (a) $f_1(r, \boldsymbol{\beta}) \propto drd\boldsymbol{\beta}/r^2$ | | | | (b) $f_2(r, \boldsymbol{\beta}) \propto drd\boldsymbol{\beta}/(1 + r)^2$ | | | |
|--|--------|-------|------------------|--|--------|-------|------------------|
| | P.M. | P.SD. | 95% P.I. | | P.M. | S.D. | 95% P.I. |
| B | 0.750 | 0.170 | (0.379, 0.990) | B | 0.752 | 0.169 | (0.375, 0.989) |
| β_1 | -1.197 | 0.131 | (-1.458, -0.936) | β_1 | -1.199 | 0.133 | (-1.458, -0.939) |
| β_2 | 0.386 | 0.193 | (0.004, 0.760) | β_2 | 0.391 | 0.190 | (0.012, 0.765) |
| (c) $f_3(r, \boldsymbol{\beta}) \propto drd\boldsymbol{\beta}/r^{1.5}$ | | | | (d) $f_4(r, \boldsymbol{\beta}) \propto drd\boldsymbol{\beta}/(1 + r)^{1.5}$ | | | |
| | P.M. | S.D. | 95% P.I. | | P.M. | S.D. | 95% P.I. |
| B | 0.862 | 0.144 | (0.495, 1.000) | B | 0.861 | 0.144 | (0.497, 1.000) |
| β_1 | -1.204 | 0.124 | (-1.450, -0.960) | β_1 | -1.198 | 0.126 | (-1.453, -0.951) |
| β_2 | 0.392 | 0.176 | (0.046, 0.743) | β_2 | 0.388 | 0.180 | (0.027, 0.736) |

and $dr/(1+r)^2$, and about 0.861 for $dr/r^{1.5}$ and $dr/(1+r)^{1.5}$.

We conduct the FMC to evaluate whether their 95% Bayesian interval estimates for random effects meet the nominal 95% confidence level. We fix the generative values of (r, β_1, β_2) at (189, -1.197, 0.386), which are posterior median for r and posterior means for β_1 and β_2 obtained under $d\boldsymbol{\beta}dr/r^2$. We simulate 3,000 ($= N_{\text{sim}}$) mock data sets, fitting the Bayesian BBLR models with four different joint HPDFs on each simulated data set. The conditions for posterior propriety are the same for all four joint HPDFs; the $k_y \times 2$ covariate matrix of interior groups (players) is of full rank 2. The probability of observing an infeasible data set is negligible (6×10^{-90}), and all 3,000 mock data sets met the condition for posterior propriety. The posterior interval of each random effect is based on 5,000 ($= N$) posterior samples obtained via the A-R method. We use the RB unbiased coverage estimator to estimate the unknown true coverage rates of random effects, i.e., $C_{r, \beta_1, \beta_2}(p_j)$ for $j = 1, 2, \dots, 18$.

In Figure 1.4, we display the results of the FMC obtained with four different HPDFs. Each plot shows 18 RB coverage estimates denoted by circles. The standard error of coverage estimates are too small to be displayed (average is 0.0002).

All four models produce 95% posterior interval estimates that meet the nominal 95% confidence level. The coverage estimates obtained with $f_1(r, \boldsymbol{\beta})$ and $f_2(r, \boldsymbol{\beta})$ (the first row of Figure 1.4) achieve the confidence level more conservatively than those obtained with $f_3(r, \boldsymbol{\beta})$ and $f_4(r, \boldsymbol{\beta})$ (the second row of Figure 1.4). The result is consistent to the previous sensitivity analysis because the length of posterior interval becomes shorter if a model produces a larger r (a larger shrinkage) a posteriori; the HPDF with exponent 1.5 tends to allow larger r a posteriori than that with exponent 2.

However, there is a limitation in this coverage statement because it is based only on a single set of generative values. Obtaining coverage estimates over all parameter values seems impossible as the parameters are continuous. We try a range of r values so that we can at least have an idea about the tendency of coverage rates, while fixing the generative values

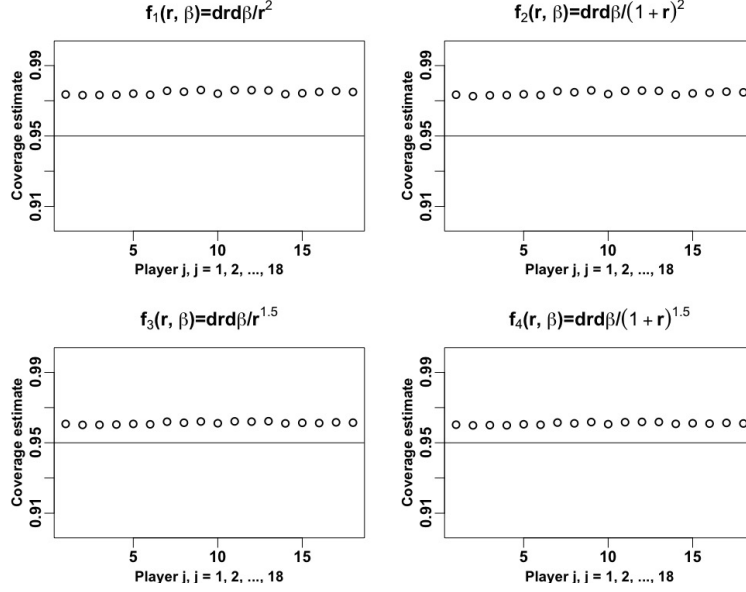


Figure 1.4: Each plot shows 18 RB coverage rate estimates based on 3,000 mock data sets generated by the values of hyper-parameters $(r, \beta_1, \beta_2) = (189, -1.197, 0.386)$.

of β_1 and β_2 .

We select equally-spaced 10 values of shrinkages inclusively between 0.05 and 0.95 and revert them to 10 generative values of r to see the effect of increasing generative values of r on coverage rate estimates ($r = 45B/(1 - B) = 2, 8, 15, 24, 37, 55, 84, 135, 255, \text{ and } 855$). The generative values of β_1 and β_2 are $(-1.179, 0.386)$, the same as before. We simulate 3,000 ($= N_{\text{sim}}$) mock data sets for each generative triple values of (r, β_1, β_2) . We fit the Bayesian BBLR models equipped with four different joint HPDFs on every mock data set. The 95% posterior intervals of random effects are based on 5,000 posterior samples of each random effect obtained via the A-R method. Because the coverage rate estimates in each plot of Figure 1.4 are almost indifferent to players, we simplify the setting by assuming $C_{r, \beta_1, \beta_2}(p_j)$ is the same as C_{r, β_1, β_2} for all random effects and use the RB overall unbiased coverage estimator.

The RB overall coverage rate estimates in Figure 1.5 show different patterns between the first and second rows. The standard errors of coverage estimates are too small to be displayed (average is 0.00014). The HPDFs on the first row, $f_1(r, \beta)$ and $f_2(r, \beta)$, produce 95%

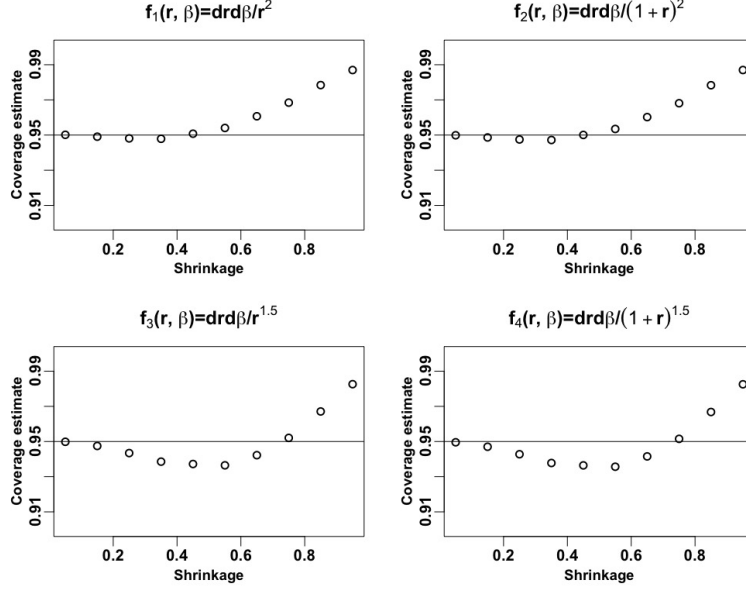


Figure 1.5: Each plot shows the 10 RB overall coverage rate estimates for $C_{r_i, \beta_1, \beta_2}$, $i = 1, 2, \dots, 10$ on corresponding 10 shrinkage factors, i.e., $(B_1, B_2, B_3, \dots, B_{10}) = (0.05, 0.15, 0.25, \dots, 0.95)$, where $B_i = r_i/(r_i + 45)$ for $i = 1, 2, \dots, 10$.

posterior intervals for random effects that slightly under-cover the nominal 95% confidence level over the range of shrinkage values between 0.15 and 0.35. On the other hand, the HPDFs on the second row, $f_3(r, \beta)$ and $f_4(r, \beta)$, produce the 95% posterior intervals that have noticeable under-coverages on the wider range of shrinkage factors between 0.15 and 0.65. It indicates that the HPDF on the first row, $f_1(r, \beta)$ and $f_2(r, \beta)$, meet the spirit of 95% confidence level better.

1.7 Conclusion

The Beta-Binomial-Logit (BBL) model accounts for the overdispersion in the Binomial data obtained from several independent groups with their covariate information considered. From a Bayesian perspective, we derive data-dependent necessary and sufficient conditions for posterior propriety of the Bayesian BBL model equipped with a joint hyper-prior density, $g(\beta)d\beta dr/(t+r)^{u+1}$, where $t \geq 0$, $u > 0$, and $g(\beta)$ can be any proper density or an improper flat density. This joint hyper-prior density encompasses those used in the literature.

Using two numerical illustrations, we look into posterior propriety and posterior properties, suggesting conservative and diffuse choices of proper hyper-prior densities be used when the posterior is improper due to improper hyper-prior probability density functions.

In the baseball example, we select four hyper-prior density functions for r and check their operating characteristics via a repeated sampling coverage evaluation, which we call *frequency method checking*. For this work, we use the Rao-Blackwellized unbiased coverage estimator to estimate unknown coverage rates of random effects and implemented an acceptance-rejection method to sample all the unknown parameters from their joint posterior distribution. It turns out that the density functions for r whose exponent is equal to 2, i.e., dr/r^2 (analog to Stein's harmonic prior) and $dr/(1+r)^2$ (uniform shrinkage prior), produce more conservative coverage rate estimates, meeting the nominal confidence level better over a wider range of generative true shrinkage values than those with exponent 1.5.

There are several opportunities to build upon our work. First of all, it is not clear whether the necessary and sufficient conditions specified in Figure 1.1 hold for other link functions, e.g., a complementary log-log link function; a probit link function is not appropriate for a BBL model because it is defined on binary data ($n_j = 1$) not on aggregate data ($n_j \geq 2$). As for frequency coverage properties, the data-dependent conditions for posterior propriety make it hard to evaluate these properties because some models with improper hyper-prior distributions do not define a frequency procedure for all possible data sets; the resulting posterior can be improper for some data sets. Thus, in a repeated sampling simulation, we may evaluate frequency properties given only the simulated data sets that achieve posterior propriety. If the probability of generating the data sets that lead to an improper posterior is negligible, this frequency evaluation procedure will be justified. We leave these for our future research.

Chapter 2

Bayesian Estimates of Astronomical Time Delays between Gravitationally Lensed Stochastic Light Curves

2.1 Introduction

Quasars are highly luminous astronomical sources in the distant Universe. The path that light takes from a quasar to Earth can be altered by the gravitational field of a massive intervening galaxy which thus acts as a lens, bending the trajectory of the emitted light; see the first panel of Figure 2.1. If the gravitational field of the galaxy is a strong gravitational lens, multiple images of the quasar can appear in slightly different locations in the sky, from the perspective of an observer on Earth, an effect known as strong gravitational lensing (Schneider et al., 1992, 2006). In this case, there are typically two or four replicate images, referred to as doubly- or quadruply-lensed quasars.

The light rays forming each of these gravitationally lensed quasar images take different routes from the quasar to Earth. Since both the lengths of the pathways and the gravitational potentials they traverse differ, the resulting multiple images are subject to differing lensing

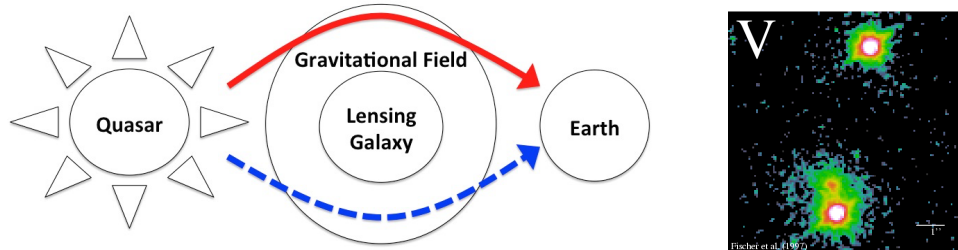


Figure 2.1: The gravitational field of an intervening galaxy acts as a lens deflecting two light rays of a quasar image towards the Earth as shown in the left panel. The arrival times can differ owing to the different lengths of pathways and different gravitational potentials they pass through. An optical V-band image of the doubly-lensed quasar Q0957+561 obtained with the Canada France Hawaii telescope (Fischer et al., 1997; Munoz et al., 1998) (<https://www.cfa.harvard.edu/castles>) appears in the right panel. The two bright sources at the top and bottom are the lensed images of the quasar, and the small red point towards the top-left of the lower quasar image is the lensing galaxy.

magnifications and their light rays arrive at the observer at different times. Because of this, any fluctuations in the source brightness are observed in each image at different times. From a statistical perspective, we can construct a time series of the brightness of each image, known as a *light curve*. Features in these light curves appear to be shifted in time and these shifts are called *time delays*.

Obtaining accurate time delay estimates is important in cosmology because they can be used to address fundamental questions regarding the origin and evolution of the Universe. For instance, Refsdal (1964) suggested using time delay estimates to constrain the Hubble constant H_o , the current expansion rate of the Universe; given a model for the mass distribution and gravitational potential of the lensing galaxy, the time delay between multiple images of the lensed quasar is inversely proportional to H_o (Blandford and Narayan, 1992; Suyu et al., 2013). Also, Linder (2011) showed that an accurate time delay estimate could substantially constrain cosmological parameters in the equation of state of dark energy characterizing the accelerated expansion of the Universe.

The upcoming large-scale astronomical survey conducted with the Large Synoptic Survey Telescope (LSST, LSST Science Collaboration, 2009), will monitor thousands of gravitationally lensed quasars beginning in 2022. The LSST is the top-ranked ground-based telescope

project in the 2010 Astrophysics Decadal Survey, and will produce extensive high-cadence time series observations of the full sky for ten years. In preparation for the *Big Data* era of the LSST, Dobler et al. (2015) organized a blind competition called the Time Delay Challenge (TDC) which ran from October 2013 to July 2014 with the aim of improving time delay estimation methods for application to realistic observational data sets. The TDC organizers prepared thousands of simulated data sets mimicking real quasar data. We are among 13 teams who took part in the TDC, each of which analyzed the simulated data using their own methods to estimate the blinded time delays¹.

2.1.1 Data and challenges

We plot a pair of simulated light curves from a doubly-lensed quasar in Figure 2.2; the light curves are labeled as A and B . Each observation time is denoted by vertical dashed lines, at which the observer measures the brightness of each gravitationally lensed quasar image. In a real data analysis, these images would correspond to the two bright sources in the second panel of Figure 2.1. The brightness is reported on the magnitude scale, an astronomical logarithmic measure of brightness, in which smaller numbers correspond to brighter objects. The magnitudes in Figure 2.2 are presented up to an overall additive calibration constant. Since the time delay is estimated via relative comparison between fluctuations in the two light curves, it is insensitive to this overall additive constant.

For a doubly-lensed quasar, there are four variables recorded on an irregularly spaced sequence of observation times $\mathbf{t} = (t_1, t_2, \dots, t_n)^\top$; the observed magnitudes denoted by $\mathbf{x}(\mathbf{t}) = (x(t_1), x(t_2), \dots, x(t_n))^\top$ for light curve A and by $\mathbf{y}(\mathbf{t}) = (y(t_1), y(t_2), \dots, y(t_n))^\top$ for light curve B , as well as the standard deviations of the measurement errors of the two light curves $\boldsymbol{\delta}(\mathbf{t}) = (\delta(t_1), \delta(t_2), \dots, \delta(t_n))^\top$ and $\boldsymbol{\eta}(\mathbf{t}) = (\eta(t_1), \eta(t_2), \dots, \eta(t_n))^\top$, respectively. In Figure 2.2, $\mathbf{x}(\mathbf{t})$ and $\mathbf{y}(\mathbf{t})$ are represented by red squares and blue circles, and their

¹In the last stage of the TDC (called *run4* in the TDC), an earlier version of our method achieved the smallest average coefficient of variation (*precision*), the TDC target for the average error level (*accuracy*) within one standard deviation, and acceptable average squared standardized residual (χ^2) after analyzing the second highest number of data sets (f). See Liao et al. (2015) for detailed result of the TDC.

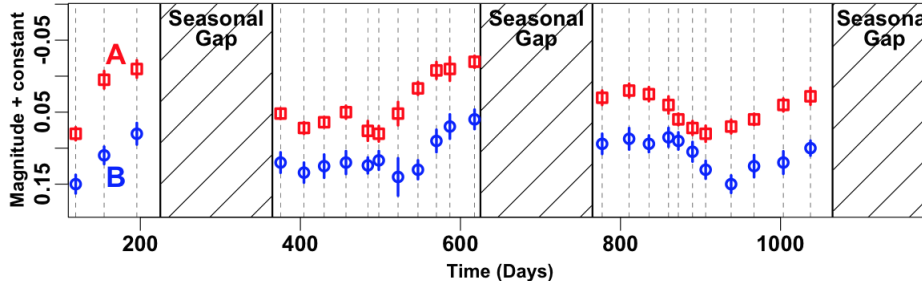


Figure 2.2: The red squares and blue circles indicate the observed magnitudes of the two simulated images at each observation time. The half lengths of vertical lines represent the standard deviations of the measurement errors. The convention in astrophysics is to plot the magnitude inversely so that smaller magnitudes (brighter object) appear on the top and larger ones (fainter object) on the bottom. The quasar magnitudes are vertically offset by an overall calibration constant, the value of which is unimportant for time delay estimation.

measurement standard errors by the half lengths of vertical lines around the magnitudes. Similarly, for a quadruply-lensed quasar, there are four light curves, each with their own measurement errors.

Since a quasar exhibits fluctuations in its brightness, it is possible to estimate time delays between different copies of those fluctuations. In Figure 2.2, for example, the bottom of the V-shaped valley of light curve *A* at around 900 days precedes that of light curve *B* by around 50 days. Other features in the light curves exhibit a similar time delay of about 50 days.

However, a number of aspects of the light curves in Figure 2.2 make accurate time delay estimation statistically challenging. First, irregular observation times are inevitable because poor weather sometimes prevents observations. Second, the motion of the Earth around the Sun causes seasonal gaps because part of the sky is not visible at night during certain months. Third, since the light of each gravitationally lensed image traverses different paths through the gravitational potential, they are subject to differing degrees of lensing magnification. Thus, the light curves often exhibit different average magnitudes. Finally, observed magnitudes are measured with error, leading to relatively larger measurement errors for fainter images.

Moreover, some quasar images exhibit additional independent extrinsic variability, an

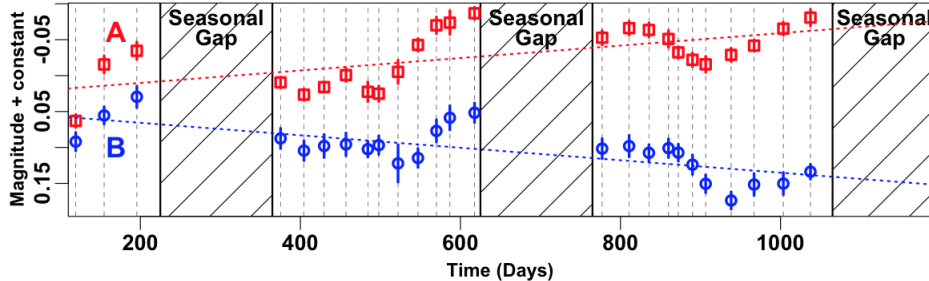


Figure 2.3: The light curves of two lensed images can have different long-term trends caused by microlensing due to stars moving within the lensing galaxy. This effect independently introduces a long-term magnification trend in each image. Here, we simulate the effect of two different long-term linear microlensing trends on the light curves in Figure 2.2. The dotted lines depict the linear microlensing trend for each image.

effect called *microlensing*. This occurs when gravitational lensing by stars moving inside the lensing galaxy independently introduces brightness magnification variations into each path of light, in addition to the overall magnifications caused by strong lensing of the galaxy (Chang and Refsdal, 1979; Tewes et al., 2013). If the timescale of the microlensing variability is much larger than that of the intrinsic quasar variability, the individual light curves may exhibit different long-term trends that are not related to the intrinsic variability of the source. As an illustration, we plot the same simulated light curves *A* and *B* with different added linear trends to simulate the effect of microlensing in Figure 2.3.

2.1.2 Other time delay estimation methods

Grid-based searches for time delay estimates are classic in this field. One-dimensional grid methods estimate the time delay, Δ_{AB} ,² between light curves *A* and *B* by minimizing the χ^2 distance or by maximizing the cross-correlation between two light curves, $\mathbf{x}(t)$ and $\mathbf{y}(t + \Delta_{AB})$, on a grid of values of Δ_{AB} (Fassnacht et al., 1999). Both techniques require an interpolation scheme. The dispersion method (Pelt et al., 1994) combines two light curves by shifting one of them in time and magnitude by Δ_{AB} and β_0 , respectively. This is called

²A positive value of Δ_{AB} indicates that features in light curve *A* appear before they appear in light curve *B*.

the *curve-shifting* assumption. The method estimates Δ_{AB} and β_0 on a two dimensional grid by minimizing the sum of squared differences between consecutive pairs of magnitudes on the combined curve. A bootstrapping method is used to produce standard errors of the time delay estimates. These methods account only for the intrinsic variability of a quasar. (When it is clear from the context, we suppress the subscript on Δ_{AB} and simply use Δ .)

Model-based methods have also been proposed in past to avoid the computational burden of evaluating the fit on a fine grid. For example, Tewes et al. (2013) model the intrinsic and extrinsic variabilities of light curves using high-order and low-order splines, respectively. They obtain the least square estimate of Δ by iterating a two-step fitting routine in which splines are first fit given Δ and then Δ is optimized given the model fit. They also use a parametric bootstrapping for the standard error of the time delay estimate.

Harva and Raychaudhury (2006, hereafter H&R) introduced the first fully Bayesian approach, though they do not account for microlensing. They assume each observed light curve is generated by an unobserved underlying process. One of the latent processes is assumed to be a shifted and scaled version of the other, with the time and magnitude shifts and the magnitude scale treated as unknown parameters. They use a collapsed Gibbs-type sampler for model fitting, with the latent process integrated out of the target posterior distribution. Unlike other existing methods this approach unifies parameter estimation and uncertainty quantification into a single coherent analysis based on the posterior distribution of Δ .

2.1.3 Our Bayesian and profile likelihood approaches

The TDC motivated us to improve on H&R’s fully Bayesian model by taking advantage of modeling and computational advances made since H&R’s 2006 proposal. Specifically, we adopt an Ornstein-Uhlenbeck (O-U) process (Uhlenbeck and Ornstein, 1930) to model the latent light curve. The O-U process has been empirically shown to describe the stochastic variability of quasar data well (Kelly et al., 2009; Kozłowski et al., 2010; MacLeod et al., 2010; Zu et al., 2013). We address the effect of microlensing by incorporating a polynomial

regression on time into the model. We specify scientifically motivated prior distributions and conduct a set of systematic sensitivity analyses. A Metropolis-Hastings (M-H) within Gibbs sampler (Tierney, 1994b) is used to take advantage of the latent process rather than integrating it out as did H&R. We improve the convergence rate of our MCMC (Markov chain Monte Carlo) sampler by using an ancillarity-sufficiency interweaving strategy (Yu and Meng, 2011) and adaptive MCMC (Brooks et al., 2011).

To complement the Bayesian method, we introduce a simple profile likelihood approach that allows us to remove nuisance parameters and focus on Δ (e.g., Davison, 2003). We show that the profile likelihood function of Δ is approximately proportional to the marginal posterior distribution of Δ when a Jeffreys' prior is used for the nuisance parameters (Berger et al., 1999), see Appendix B.4. For the problems we investigate the profile likelihood is nearly identical to the marginal posterior distribution, validating both methods.

Our time delay estimation strategy combines these two complementary approaches. We first obtain the profile likelihood of Δ , which is simple to compute. A more principled fully Bayesian analysis focuses on the dominant mode identified by the profile likelihood and provides joint inference for the time delay and other model parameters via the joint posterior distribution.

The rest of this chapter is organized as follows. We describe our Bayesian model in Section 2.2 and the MCMC sampler that we use to fit it in Section 2.3. In Section 2.4, we introduce the profile likelihood approach. We then specify our estimation strategy and illustrate it via a set of numerical examples in Section 2.5. An R package, `timedelay`, that implements the Bayesian and profile likelihood methods is publicly available at CRAN (<https://cran.r-project.org/package=timedelay>).

2.2 A fully Bayesian model for time delay estimation

2.2.1 Latent time series

We assume that each time-delayed light curve is generated from a latent curve representing the true source magnitude in continuous time. For example, the solid red and dashed blue curves in Figure 2.4 are the latent light curves and are denoted by $\mathbf{X} = \{X(t), t \in \mathbf{R}\}$ and $\mathbf{Y} = \{Y(t), t \in \mathbf{R}\}$, respectively, where $X(t)$ and $Y(t)$ are unobserved true magnitudes at time t . We use the vector notation $\mathbf{X}(\mathbf{t}) = (X(t_1), X(t_2), \dots, X(t_n))^T$ and $\mathbf{Y}(\mathbf{t}) = (Y(t_1), Y(t_2), \dots, Y(t_n))^T$ to denote the n magnitudes of each latent light curve at the irregularly-spaced observation times \mathbf{t} .

A curve-shifted model (Pelt et al., 1994; Kochanek et al., 2006) assumes that one of the latent light curves is a shifted version of the other, that is

$$Y(t) = X(t - \Delta) + \beta_0, \quad (2.1)$$

where Δ is a shift in time and β_0 is a magnitude offset. For example, we generated the latent curves in Figure 2.4 under the model in (2.1). Thus the two latent curves exactly overlap if the solid red curve is shifted by Δ days and by β_0 magnitude units. The key advantage of this model is that a single latent light curve, here \mathbf{X} , is sufficient to represent the true magnitude time series of the two (or more) lensed images. This model is a special case of H&R’s scaled

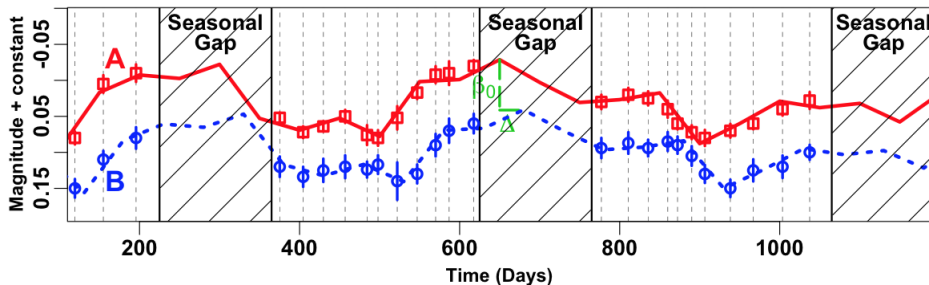


Figure 2.4: The dashed blue and solid red curves represent the latent continuous time light curves and are superimposed on Figure 2.2. The curve-shifted model in (2.1) specifies that the blue dashed curve is a shifted version of the red curve by Δ ($=70$) days in time and by β_0 ($=0.07$) in magnitude.

curve-shifted model, $Y(t) = sX(t - \Delta) + \beta_0$, where s is a magnitude scale change, mentioned at the end of Section 2.1.2. Setting $s = 1$ is reasonable because gravitational lensing only deflects the source light and magnifies it, i.e., multiplies the source flux. Because magnitude is on the \log_{10} scale of source flux, we expect an additive offset, i.e., β_0 , rather than a scale change. Since the curve-shifted model reflects gravitational lensing well, it is appropriate for estimating Δ , at least in the absence of microlensing.

Microlensing causes additional long-term extrinsic variability unrelated to the intrinsic quasar variability driving the dynamics of \mathbf{X} . Thus, the curve-shifted model is not appropriate in the presence of microlensing. To account for microlensing, we assume that one of the latent light curves is a time-shifted version of the other, but with an additional m -order polynomial regression on $t - \Delta$, that is

$$Y(t) = X(t - \Delta) + \mathbf{w}_m^\top(t - \Delta)\boldsymbol{\beta}, \quad (2.2)$$

where $\mathbf{w}_m(t - \Delta) \equiv (1, t - \Delta, (t - \Delta)^2, \dots, (t - \Delta)^m)^\top$ is a covariate vector of length $m + 1$, and $\boldsymbol{\beta} \equiv (\beta_0, \beta_1, \beta_2, \dots, \beta_m)^\top$ is a vector of regression coefficients. The polynomial regression term in (3.15) accounts for the difference in the microlensing trends of the two light curves, i.e., the difference between the long-term trends of $Y(t)$ and $X(t - \Delta)$. The microlensing model in (3.15) reduces to a curve-shifted model in (2.1) if $\beta_1 = \beta_2 = \dots = \beta_m = 0$.

The best choice for the order of the polynomial regression depends on the extent of microlensing, and this varies from quasar to quasar. We set $m = 3$ as a default because the third order polynomial regression has been successfully applied to model lensed quasars (Kochanek et al., 2006; Morgan et al., 2012; Courbin et al., 2013). If we find evidence via the profile likelihood that a third order polynomial regression is not sufficient to reduce the effect of microlensing (see Section 2.5.1 for details), we can impose a reasonable upper bound of m by running preliminary regression on the observed light curves, and comparing the fits.

2.2.2 Distribution of the observed data

Observing the gravitationally-lensed images with a telescope, an astronomer measures the magnitude in each image, $x(t_j)$ and $y(t_j)$, along with the standard deviations, $\delta(t_j)$ and $\eta(t_j)$, at time t_j , $j = 1, 2, \dots, n$. We assume that these measurements have independent Gaussian errors centered at the latent magnitudes $X(t_j)$ and $Y(t_j)$, i.e.,

$$x(t_j) | X(t_j) \stackrel{\text{indep.}}{\sim} \text{N}[X(t_j), \delta^2(t_j)], \quad (2.3)$$

$$y(t_j) | Y(t_j) \stackrel{\text{indep.}}{\sim} \text{N}[Y(t_j), \eta^2(t_j)], \quad (2.4)$$

where $\text{N}[M, V]$ is a Gaussian distribution with mean M and variance V , and $\mathbf{x}(\mathbf{t})$ and $\mathbf{y}(\mathbf{t})$ are independent given their true magnitudes. Using the model in (3.15), we can express (3.16) as

$$y(t_j) | X(t_j - \Delta), \Delta, \boldsymbol{\beta} \stackrel{\text{indep.}}{\sim} \text{N}[X(t_j - \Delta) + \mathbf{w}_m^\top(t_j - \Delta)\boldsymbol{\beta}, \eta^2(t_j)]. \quad (2.5)$$

Given Δ , we define $\mathbf{t}^\Delta = (t_1^\Delta, t_2^\Delta, \dots, t_{2n}^\Delta)^\top$ as the sorted vector of $2n$ times among the n observation times, \mathbf{t} , and the n time-delay-shifted observation times, $\mathbf{t} - \Delta$. Also, $\mathbf{X}(\mathbf{t}^\Delta) = (X(t_1^\Delta), X(t_2^\Delta), \dots, X(t_{2n}^\Delta))^\top$ is the vector of $2n$ latent magnitudes at the times in \mathbf{t}^Δ . The joint density function of the observed data given $\mathbf{X}(\mathbf{t}^\Delta)$, Δ , and $\boldsymbol{\beta}$ is

$$p(\mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}) | \mathbf{X}(\mathbf{t}^\Delta), \Delta, \boldsymbol{\beta}) = \prod_{j=1}^n p(x(t_j) | X(t_j)) \times p(y(t_j) | X(t_j - \Delta), \Delta, \boldsymbol{\beta}), \quad (2.6)$$

where the two distributions in the product are given in (2.3) and (2.5).

2.2.3 Distribution of the latent magnitudes

We assume the latent continuous time light curve, \mathbf{X} , is a realization of an O-U process (Uhlenbeck and Ornstein, 1930) as proposed in Kelly et al. (2009, hereafter KBS). The stochastic differential equation,

$$dX(t) = -\frac{1}{\tau}(X(t) - \mu)dt + \sigma dB(t), \quad (2.7)$$

defines the O-U process, where μ and σ are on the magnitude scale and govern the overall mean and short-term variability of the underlying process, τ is a timescale (in days) for the process to revert to the long-term mean μ , $\{B(t), t \geq 0\}$ is a standard Brownian motion, and $dB(t)$ is an interval of the Brownian motion, whose distribution is Gaussian with mean zero and variance dt . We denote the three O-U parameters by $\boldsymbol{\theta} = (\mu, \sigma^2, \tau)^\top$.

KBS empirically demonstrated that the power spectrum of the O-U process is consistent with the mean power spectrum of 55 well-sampled quasar light curves at a specific frequency range with timescales shorter than τ . KBS also investigated the associations between model parameters and the physical properties of quasars. For example, τ has a positive correlation with black hole mass, which is consistent with previous astrophysical studies. Kozłowski et al. (2010) and MacLeod et al. (2010) were concerned about a possible selection bias in the sample of quasars used in KBS and thus they analyzed thousands of light curves. Kozłowski et al. (2010) found further support for the O-U process in their analyses of about 2,700 quasars obtained from the Optical Gravitational Lensing Experiment (OGLE, Kozłowski and Kochanek, 2009). They showed that the distribution of the goodness of fit statistic obtained by fitting the O-U process to their light curves was consistent to the expected distribution of the statistic under the assumption that the light curve variation was stochastic. MacLeod et al. (2010) further verified the argument about the correlations between model parameters and physical properties in KBS by analyzing about 9,000 quasars obtained from the Sloan Digital Sky Survey (Berk et al., 2004). Zu et al. (2013) also supported the O-U process by comparing it to the Gaussian process with three different covariance functions in fitting about 200 OGLE light curves. Their numerical results based on the F -test and Bayesian information criterion supported the O-U process. These studies popularized the O-U process among astrophysicists to the extent that the TDC simulated its quasar light curves under an O-U process (Dobler et al., 2015). The earlier approach of H&R (2006) preceded these more recent advances in astrophysical and statistical modeling of quasars.

The solution of the stochastic differential equation in (2.7) provides the sampling distri-

bution for the time-sorted latent magnitudes $\mathbf{X}(t^\Delta)$ via its Markovian property. Specifically,

$$\begin{aligned} X(t_1^\Delta) \mid \Delta, \boldsymbol{\theta} &\sim \text{N} \left[\mu, \frac{\tau\sigma^2}{2} \right], \text{ and for } j = 2, 3, \dots, 2n, \\ X(t_j^\Delta) \mid X(t_{j-1}^\Delta), \Delta, \boldsymbol{\theta} &\sim \text{N} \left[\mu + a_j(X(t_{j-1}^\Delta) - \mu), \frac{\tau\sigma^2}{2}(1 - a_j^2) \right], \end{aligned} \quad (2.8)$$

where $a_j \equiv \exp(-(t_j^\Delta - t_{j-1}^\Delta)/\tau)$ is a shrinkage factor that depends on the observational cadence and τ . If two adjacent latent magnitudes are close in time, i.e., $t_j^\Delta - t_{j-1}^\Delta$ is small, a_j is close to one and $X(t_j^\Delta)$ borrows more information or shrinks more towards the previous latent magnitude, $X(t_{j-1}^\Delta)$, and exhibits less uncertainty. On the other hand, if neighboring latent magnitudes are distant in time, e.g., due to a seasonal gap, a_j is close to zero, and $X(t_j^\Delta)$ borrows little information from the distant value $X(t_{j-1}^\Delta)$ and instead approaches the overall mean μ . This is known as *mean reversion* property of the O-U process.

The joint density function of the $2n$ latent magnitudes is

$$p(\mathbf{X}(t^\Delta) \mid \Delta, \boldsymbol{\theta}) = p(X(t_1^\Delta) \mid \Delta, \boldsymbol{\theta}) \times \prod_{j=2}^{2n} p(X(t_j^\Delta) \mid X(t_{j-1}^\Delta), \Delta, \boldsymbol{\theta}), \quad (2.9)$$

where the distributions on the right-hand side are given in (2.8).

2.2.4 Prior distributions for the time delay and the magnitude offset

We adopt the proper prior distributions, for Δ and $\boldsymbol{\beta}$,

$$p(\Delta, \boldsymbol{\beta}) = p(\Delta)p(\boldsymbol{\beta}) \propto I_{\{u_1 \leq \Delta \leq u_2\}} \times \text{N}_{m+1}(\boldsymbol{\beta} \mid \mathbf{0}, 10^5 \times I_{m+1}), \quad (2.10)$$

where $I_{\{D\}}$ is the indicator function of D , $\text{N}_{m+1}(\boldsymbol{\beta} \mid \mathbf{0}, I_{m+1})$ is an $m+1$ dimensional Gaussian density evaluated at $\boldsymbol{\beta}$ whose mean is $\mathbf{0}$, a vector of zeros with length $m+1$, and variance-covariance matrix is $10^5 \times I_{m+1}$, with an $m+1$ dimensional identity matrix I_{m+1} . We put a diffuse Gaussian prior on $\boldsymbol{\beta}$ to minimize impact on the posterior inference and to ensure posterior propriety.

The range of the uniform prior distribution on Δ , $[u_1, u_2]$, reflects the range of interest. One choice is the *entire feasible range* (or feasible range) of Δ , $[t_1 - t_n, t_n - t_1]$, where there is at least one data point that overlaps between the two light curves. Outside of this range, the two light curves do not overlap and the data cannot identify Δ . (H&R used a diffuse Gaussian prior distribution on Δ that was defined outside its feasible range.)

In some cases, information about the likely range of Δ is available from previous analyses or possibly from astrophysical probes. To find the likely range of Δ , we can also use a physical model for the mass and gravitational potential of the lens, as well as the redshifts (an astronomical measure of distance) and relative spatial locations of a quasar and lens.

2.2.5 Prior distributions for the parameters in the O-U process

Considering both scientific knowledge and the dynamics of the O-U process, we put a uniform distribution on the O-U mean μ , an independent inverse-Gamma (IG) distribution, $\text{IG}(1, b_\sigma)$, on its short-term variance σ^2 , and an independent $\text{IG}(1, 1)$ distribution on its timescale τ , i.e.,

$$p(\mu, \sigma^2, \tau) = p(\mu)p(\sigma^2)p(\tau) \propto \frac{\exp(-b_\sigma/\sigma^2)}{(\sigma^2)^2} \cdot \frac{\exp(-1/\tau)}{\tau^2} \quad (2.11)$$

$$\times I_{\{-30 \leq \mu \leq 30\}} \cdot I_{\{\sigma^2 > 0\}} \cdot I_{\{\tau > 0\}}.$$

Here the uniform distribution on μ encompasses a magnitude range from that of the Sun (magnitude = -26.74) to that of the faintest object visible with the Hubble Space Telescope (magnitude = 30). The IG distributions on τ and σ^2 set their soft lower bounds³ to focus on practical solutions in which Δ can be constrained. For example, in the limits when τ is much less than the observation cadence or when σ^2 is much smaller than the measurement variance divided by the cadence, the discrete observations of the continuous latent light curve appear as serially uncorrelated white noise sequence. In these limiting cases it is impossible

³Because the density function of $\text{IG}(a, b)$ decreases exponentially from its mode, $b/(a + 1)$, toward zero and geometrically decreases with a power of $a + 1$ towards infinity, it is relatively unlikely for the random variable to take on values smaller than its mode.

to estimate Δ by matching serially correlated fluctuation patterns. The soft lower bounds for τ and σ^2 discount these limiting cases, and allow us to focus on the relevant parameter space in which we expect time delay estimation to be feasible.

The relationship between the IG and scaled inverse- χ^2 distributions allows us to interpret the shape parameter of the IG as half the number of directly observed pseudo realizations of the O-U process that would carry equivalent information as the prior distribution. (See, e.g., Gelman et al. (2013) for a discussion of the pseudo observation interpretation of prior distributions.) Thus, we set the shape parameter of the IG prior on τ to one; this corresponds to two pseudo observations and can be interpreted as an indication that the prior distribution is relatively weak. It is practical to set the scale parameter of the IG prior for τ to one; the resulting soft lower bound on τ is 0.5 day and is smaller than all estimates of τ in MacLeod et al. (2010), who analyzed 9,275 quasars.

For the IG prior distribution of σ^2 , we set the shape parameter to one and the scale parameter⁴ to $(\text{Mean measurement standard error})^2 / (\text{Median cadence})$, i.e.,

$$b_\sigma = \frac{[\{\sum_{j=1}^n \delta(t_j) + \sum_{j=1}^n \eta(t_j)\}/2n]^2}{\text{Median}(t_2 - t_1, t_3 - t_2, \dots, t_n - t_{n-1})}. \quad (2.12)$$

This scale parameter enables us to search for solutions for which we can constrain Δ by avoiding the above limiting cases. Another viable choice for the scale parameter is $b_\sigma = 2/10^7$ because all estimates of σ^2 in MacLeod et al. (2010) are larger than this value. Sensitivity analyses for the choice of prior distributions of τ and σ^2 appear in Appendix B.5.

2.3 Metropolis-Hastings within Gibbs sampler

Our overall hierarchical model is specified via the observation model in (2.3) and (2.5), the O-U process for the latent light curve in (2.8), and the prior distributions given in (2.10) and (2.11). Our first approach to model fitting uses a Gibbs-type sampler to explore the resulting full posterior distribution. Instead of integrating out the latent magnitudes and

⁴The physical unit of b_σ is magnitude squared per day, hereafter mag² per day.

using a collapsed sampler as H&R did, we treat $\mathbf{X}(\mathbf{t}^\Delta)$ as latent variables, alternatively updating $\mathbf{X}(\mathbf{t}^\Delta)$ and the other model parameters. (We could formulate our approach as data augmentation with $\mathbf{X}(\mathbf{t}^\Delta)$ as the missing data, see van Dyk and Meng (2001).)

Specifically, we use a Metropolis-Hastings within Gibbs (MHwG) sampler (Tierney, 1994b) that iteratively samples five complete conditional distributions of the full joint posterior density, $p(\mathbf{X}(\mathbf{t}^\Delta), \Delta, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$, proportional to the product of densities of observed and latent data in (2.6) and (2.9) and prior densities in (2.10) and (2.11). Iteration l of our sampler is composed of five steps.

$$\text{Step 1: Sample } (\mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}}), \Delta^{(l)}) \sim p(\mathbf{X}(\mathbf{t}^\Delta), \Delta \mid \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}) \quad (2.13)$$

$$= p(\mathbf{X}(\mathbf{t}^\Delta) \mid \Delta, \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}) \times p(\Delta \mid \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}) \text{ by M-H}$$

$$\text{Step 2: Sample } \boldsymbol{\beta}^{(l)} \sim p(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(l-1)}, \mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}}), \Delta^{(l)}) \quad (2.14)$$

$$\text{Step 3: Sample } \mu^{(l)} \sim p(\mu \mid (\sigma^2)^{(l-1)}, \tau^{(l-1)}, \mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}}), \Delta^{(l)}, \boldsymbol{\beta}^{(l)}) \quad (2.15)$$

$$\text{Step 4: Sample } (\sigma^2)^{(l)} \sim p(\sigma^2 \mid \tau^{(l-1)}, \mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}}), \Delta^{(l)}, \boldsymbol{\beta}^{(l)}, \mu^{(l)}) \quad (2.16)$$

$$\text{Step 5: Sample } \tau^{(l)} \sim p(\tau \mid \mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}}), \Delta^{(l)}, \boldsymbol{\beta}^{(l)}, \mu^{(l)}, (\sigma^2)^{(l)}) \text{ by M-H,} \quad (2.17)$$

where we suppress conditioning on $\mathbf{x}(\mathbf{t})$ and $\mathbf{y}(\mathbf{t})$ in all five steps. The conditional distributions in (2.14), (2.15), and (2.16), are standard families that can be sampled directly, whereas those in (2.13) and (2.17) require M-H updates. We use the factorization in (2.13) to construct a joint proposal, $(\tilde{\mathbf{X}}(\mathbf{t}^{\tilde{\Delta}}), \tilde{\Delta})$, for $(\mathbf{X}(\mathbf{t}^\Delta), \Delta)$ and calculate its acceptance probability. First, $\tilde{\Delta}$ is proposed from the Gaussian density $\text{N}(\Delta^{(l-1)}, \psi^2)$, where ψ is a proposal scale and is set to produce a reasonable acceptance rate. Given $\tilde{\Delta}$, we propose $\tilde{\mathbf{X}}(\mathbf{t}^{\tilde{\Delta}}) \sim p(\mathbf{X}(\mathbf{t}^{\tilde{\Delta}}) \mid \tilde{\Delta}, \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}, \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$; this is a Gaussian distribution and is specified in Appendix B.1. Because the proposal for Δ and that for $\mathbf{X}(\mathbf{t}^\Delta)$ given Δ are symmetric, $(\tilde{\mathbf{X}}(\mathbf{t}^{\tilde{\Delta}}), \tilde{\Delta})$ is accepted with a probability $\min(1, r)$, where

$$r = \frac{p(\tilde{\Delta} \mid \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}, \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))}{p(\Delta^{(l-1)} \mid \boldsymbol{\beta}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}, \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))}. \quad (2.18)$$

Details of the marginalized density $p(\Delta \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$ in (2.18) appear in Appendix B.2 and details of Step 2–5 appear in Appendix B.3.

The direct updates for β , μ , and σ^2 are based on standard families that are not available under H&R’s collapsed approach. Thus the collapsed approach must update each of the model parameters via a Metropolis or M-H update, which can slow down the rate of convergence. (Collapsing Gibbs-type samplers, however, is known to improve their rate of convergence (Liu, 2008) if the complete conditionals can be sampled directly.) Also, the collapsed MHwG (CMHwG) sampler requires about three times more CPU time per iteration than the (non-collapsed) MHwG sampler that we propose. In Figure 2.5, we compare the autocorrelation functions (ACFs) of Δ , β_0 , μ , σ^2 , and τ obtained by the CMHwG sampler (first row) and those obtained by our MHwG sampler (second row). The sampler in the third row is discussed in Section 2.3.1. All algorithms are run using the curve-shifted model in (2.1) fit to data for quasar *Q0957+561* (Hainline et al., 2012). Except for that of β_0 , the ACFs generated with CMHwG (first row), decay slower than those obtained with MHwG (second row). The effective sample sizes per second (ESS/sec) tend to improve with MHwG over CMHwG. For example, for Δ the ESS/sec is 5.23 with CMHwG and 21.09 with MHwG. The exception is β_0 , for which ESS/sec is 6.33 with CMHwG, but only 1.74 with MHwG. In

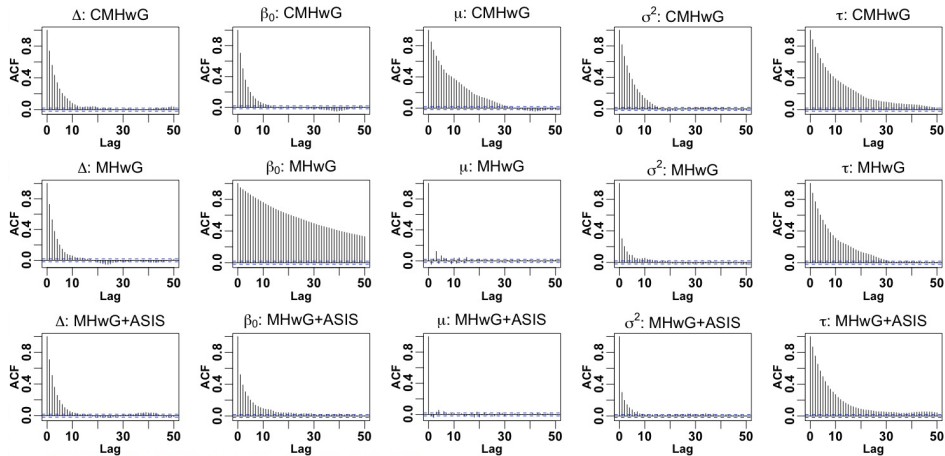


Figure 2.5: The autocorrelation functions for Δ , β_0 , μ , σ^2 , and τ (columns from left to right) based on 10,000 posterior samples after a burn-in of 10,000. Results are obtained using three different posterior samplers (CMHwG, MHwG, and MHwG + ASIS, rows from top to bottom). We use the curve-shifted model for simplicity and the data from quasar Q0957+671 analyzed in Section 2.5.3.

the following section, we discuss a way to improve the convergence rate of β_0 (β in general) for the MHwG sampler, while retaining its fast running time.

2.3.1 Ancillarity-sufficiency interweaving strategy

To improve the convergence rate of β , we adopt the ancillarity-sufficiency interweaving strategy (ASIS, Yu and Meng, 2011). ASIS interweaves trajectories of Markov chains of β obtained by two discordant parameterizations of the unknown quantities. The two parameterizations are designed so that the latent variables can be viewed as ancillary and sufficient statistics for β , respectively.

In the parameterization used up until now, $\mathbf{X}(\mathbf{t}^\Delta)$ is an *ancillary augmentation* (AA) for β in that it is an ancillary statistic for β . That is, the distribution of $\mathbf{X}(\mathbf{t}^\Delta)$ in (2.8) does not depend on β . On the other hand, a *sufficiency augmentation* (SA) for β is based on the latent variables that have sufficient information to estimate β , that is, a sufficient statistic for β . To derive a SA for β , we introduce the parameterization,

$$K(t_j^\Delta) \equiv X(t_j^\Delta) + \mathbf{w}_m^\top(t_j^\Delta)\beta \cdot I_{\mathbf{t}-\Delta}(t_j^\Delta), \text{ for } j = 1, 2, \dots, 2n, \quad (2.19)$$

where

$$I_{\mathbf{t}-\Delta}(t_j^\Delta) = \begin{cases} 1, & \text{if } t_j^\Delta \in \mathbf{t} - \Delta, \\ 0, & \text{if } t_j^\Delta \in \mathbf{t}. \end{cases} \quad (2.20)$$

This indicator is one if t_j^Δ is an element of $\mathbf{t} - \Delta = \{t_1 - \Delta, t_2 - \Delta, \dots, t_n - \Delta\}$ and zero otherwise. Using (2.19), we express the observation model in (2.3) and (2.5) as

$$x(t_j) | K(t_j) \stackrel{\text{indep.}}{\sim} \text{N}[K(t_j), \delta^2(t_j)]. \quad (2.21)$$

$$y(t_j) | K(t_j - \Delta), \Delta \stackrel{\text{indep.}}{\sim} \text{N}[K(t_j - \Delta), \eta^2(t_j)]. \quad (2.22)$$

The distributions of latent light curve in (2.8) is replaced by

$$\begin{aligned}
K(t_1^\Delta) \mid \Delta, \boldsymbol{\beta}, \boldsymbol{\theta} &\sim \text{N} \left[\mu + \mathbf{w}_m^\top(t_1^\Delta) \boldsymbol{\beta} \cdot I_{\{t-\Delta\}}(t_1^\Delta), \frac{\tau\sigma^2}{2} \right], & (2.23) \\
K(t_j^\Delta) \mid K(t_{j-1}^\Delta), \Delta, \boldsymbol{\beta}, \boldsymbol{\theta} &\sim \text{N} \left[\mu + \mathbf{w}_m^\top(t_j^\Delta) \boldsymbol{\beta} \cdot I_{\{t-\Delta\}}(t_j^\Delta) \right. \\
&\quad \left. + a_j (K(t_{j-1}^\Delta) - \mu - \mathbf{w}_m^\top(t_{j-1}^\Delta) \boldsymbol{\beta} \cdot I_{\{t-\Delta\}}(t_{j-1}^\Delta)), \frac{\tau\sigma^2}{2} (1 - a_j^2) \right].
\end{aligned}$$

Under this reparameterization of the model in terms of $\mathbf{K}(\mathbf{t}^\Delta)$, $\boldsymbol{\beta}$ appears only in (2.23), which means that $\mathbf{K}(\mathbf{t}^\Delta)$ contain sufficient information to estimate $\boldsymbol{\beta}$ and thus $\mathbf{K}(\mathbf{t}^\Delta)$ is SA for $\boldsymbol{\beta}$. Because the parameterization does not affect the prior distributions of the model parameters in (2.10) and (2.11), the full joint posterior density in terms of $\mathbf{K}(\mathbf{t}^\Delta)$, i.e., $p(\mathbf{K}(\mathbf{t}^\Delta), \Delta, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$, is proportional to the product of densities of observed and latent data, whose distributions are specified in (2.21), (2.22), and (2.23), and prior densities in (2.10) and (2.11). Consequently, the marginal posterior distribution of the model parameters, $\{\Delta, \boldsymbol{\beta}, \boldsymbol{\theta}\}$, is unchanged.

ASIS interweaves the trajectory of $\boldsymbol{\beta}$ from a sample constructed under AA and that constructed under SA. This can be accomplished by replacing Step 2 in (2.14) with the following four steps:

$$\text{Step } 2_a : \text{Sample } \boldsymbol{\beta}_{\text{AA}}^{(l)} \sim p(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(l-1)}, \mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}}), \Delta^{(l)}) \quad (2.24)$$

$$\text{Step } 2_b : \text{Set } K^{(l)}(t_j^{\Delta^{(l)}}) = X^{(l)}(t_j^{\Delta^{(l)}}) + \mathbf{w}_m^\top(t_j^{\Delta^{(l)}}) \boldsymbol{\beta}_{\text{AA}}^{(l)} I_{t-\Delta^{(l)}}(t_j^{\Delta^{(l)}}) \quad (2.25)$$

$$\text{Step } 2_c : \text{Sample } \boldsymbol{\beta}_{\text{SA}}^{(l)} \sim p(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(l-1)}, \mathbf{K}^{(l)}(\mathbf{t}^{\Delta^{(l)}}), \Delta^{(l)}) \quad (2.26)$$

$$\text{Step } 2_d : \text{Set } X^{(l)}(t_j^{\Delta^{(l)}}) = K^{(l)}(t_j^{\Delta^{(l)}}) - \mathbf{w}_m^\top(t_j^{\Delta^{(l)}}) \boldsymbol{\beta}_{\text{SA}}^{(l)} I_{t-\Delta^{(l)}}(t_j^{\Delta^{(l)}}) \quad (2.27)$$

Again, we suppress the condition on $\mathbf{x}(\mathbf{t})$ and $\mathbf{y}(\mathbf{t})$. In Step 2_c , we set $\boldsymbol{\beta}^{(l)}$ to $\boldsymbol{\beta}_{\text{SA}}^{(l)}$ sampled from its conditional posterior distribution specified in (B.11). In Step 2_d , ASIS updates the latent variables, $\mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}})$, to adjust for the inconsistency between the updates sampled in (2.15)–(2.17) that are based on the $\mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}})$ and the update $\boldsymbol{\beta}^{(l)}$ that is based on $\mathbf{K}^{(l)}(\mathbf{t}^{\Delta^{(l)}})$. Updating $\mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}})$ in (2.27) synchronizes this inconsistency and preserves the stationary distribution (Yu and Meng, 2011). The additional computational cost of ASIS

is negligible because the conditional updates in (2.24) and (2.26) include quick multivariate Gaussian sampling, see Appendix (B.7) and (B.11) for details.

ACFs of the model parameters obtained by MHwG equipped with ASIS, denoted by MHwG+ASIS, appear on the third row of Figure 2.5; the ACF of β_0 in the second column shows a noticeable improvement compared to that obtained by MHwG sampler. The ESS/sec for β_0 is 20.95 with MHwG+ASIS and 1.74 with MHwG. (The ESS/sec for Δ is 21.35 with MHwG+ASIS and 21.09 with MHwG.)

2.3.2 Adaptive MCMC

Our MHwG sampler (either with or without ASIS) requires a proposal distribution in each of its two Metropolis steps, that is, $N[\Delta^{(l-1)}, \psi^2]$ used to update $\Delta^{(l)}$ in (2.13) and $N[\log(\tau^{(l-1)}), \phi^2]$ used to update $\log(\tau^{(l)})$ in (2.17), where ψ and ϕ are the proposal scales. To avoid burdensome off-line tuning of the proposal scales, we implement an adaptive MCMC (Brooks et al., 2011) that allows automatical adjustment during the run. We implement an algorithm that updates the two proposal scales every 100 iterations, based on the most recent 100 proposals as outlined in Step 6 of Figure 2.6. The Markov chains equipped with the adaptive MCMC converge to the stationary distribution because the adjustment factors, $\exp(\pm \min(0.01, 1/\sqrt{i}))$, in Step 6 of of Figure 2.6 approach one as i goes to infinity. This condition is called diminishing adaptation condition (Roberts and Rosenthal, 2007). We set the lower and upper bounds of the acceptance rate to 0.23 and 0.44, respectively (Gelman et al., 2013).

The steps of the adaptive MHwG+ASIS sampler are specified in Figure 2.6. We describe our choice of initial values of the parameters in Section 2.5 as a part of our time delay estimation strategy.

Set $\mathbf{X}^{(0)}(\mathbf{t}^{\Delta^{(0)}})$, $\Delta^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\mu^{(0)}$, $(\sigma^2)^{(0)}$, $\tau^{(0)}$, $\psi^{(0)}$, $\phi^{(0)}$.

For $l = 1, 2, \dots$

Step 1: Sample $\Delta^{(l)}$ using a Metropolis step with proposal rule $N[\Delta^{(l-1)}, (\psi^{(l-1)})^2]$.

If a new proposal for $\Delta^{(l)}$ is accepted, then sample $\mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}})$,
or otherwise set $\mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}})$ to $\mathbf{X}^{(l-1)}(\mathbf{t}^{\Delta^{(l-1)}})$.

Step 2: (ASIS) Update $\boldsymbol{\beta}^{(l)}$ and $\mathbf{X}^{(l)}(\mathbf{t}^{\Delta^{(l)}})$ via (2.24)–(2.27).

Step 3: Sample $\mu^{(l)}$ via (2.15).

Step 4: Sample $(\sigma^2)^{(l)}$ via (2.16).

Step 5: Sample $\tau^{(l)}$ using a M-H step with proposal rule $N[\log(\tau^{(l-1)}), (\phi^{(l-1)})^2]$.

Step 6: (Adaptation) If $l \bmod 100 = 0$

if the acceptance rate of Δ in iterations $l - 99, l - 98, \dots, l > 0.44$ **then**

$\psi^{(l)} \leftarrow \psi^{(l-1)} \times \exp(\min(0.01, 1/\sqrt{(l/100)}))$

else if the acceptance rate of Δ in iterations $l - 99, l - 98, \dots, l < 0.23$ **then**

$\psi^{(l)} \leftarrow \psi^{(l-1)} \times \exp(-\min(0.01, 1/\sqrt{(l/100)}))$

end if

if the acceptance rate of τ in iterations $l - 99, l - 98, \dots, l > 0.44$ **then**

$\phi^{(l)} \leftarrow \phi^{(l-1)} \times \exp(\min(0.01, 1/\sqrt{(l/100)}))$

else if the acceptance rate of τ in iterations $l - 99, l - 98, \dots, l < 0.23$ **then**

$\phi^{(l)} \leftarrow \phi^{(l-1)} \times \exp(-\min(0.01, 1/\sqrt{(l/100)}))$

end if

Otherwise $\psi^{(l)} = \psi^{(l-1)}$ and $\phi^{(l)} = \phi^{(l-1)}$.

Figure 2.6: Steps of the adaptive MHwG+ASIS sampler.

2.4 Profile likelihood of the time delay

A profile likelihood of Δ (e.g., Davison, 2003) is a simple approximation to the marginal posterior distribution of Δ , $p(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$. It is defined as

$$L_{\text{prof}}(\Delta) \equiv \max_{\boldsymbol{\beta}, \boldsymbol{\theta}} L(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta}) = L(\Delta, \hat{\boldsymbol{\beta}}_{\Delta}, \hat{\boldsymbol{\theta}}_{\Delta}), \quad (2.28)$$

where $L(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta})$ is the likelihood function of the model parameters, that is,

$$\begin{aligned} L(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta}) &= p(\mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}) \mid \Delta, \boldsymbol{\beta}, \boldsymbol{\theta}) \\ &= \int p(\mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}) \mid \mathbf{X}(\mathbf{t}^{\Delta}), \Delta, \boldsymbol{\beta}) \times p(\mathbf{X}(\mathbf{t}^{\Delta}) \mid \Delta, \boldsymbol{\theta}) d\mathbf{X}(\mathbf{t}^{\Delta}), \end{aligned} \quad (2.29)$$

and $(\hat{\boldsymbol{\beta}}_{\Delta}, \hat{\boldsymbol{\theta}}_{\Delta})$ are the values of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ that maximize $L(\Delta, \boldsymbol{\beta}, \boldsymbol{\theta})$ for each Δ .

In regular problems the profile likelihood of a parameter, say φ , approximates its marginal posterior distribution with a uniform prior on φ . This happens, for example, if the log likelihood of the model parameters is approximately quadratic given φ under standard asymptotic

arguments. The prior distribution on the parameters other than φ is chosen in such a way as to approximately overset the variance term of the log likelihood, e.g., as happens asymptotically with the Jeffreys' prior, see Appendix D for details.

Treating $L_{\text{prof}}(\Delta)$ as an approximation to $p(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$, we evaluate $L_{\text{prof}}(\Delta)$ on a fine grid of values over the interesting range of Δ , $\{\Delta_1, \Delta_2, \dots, \Delta_w\}$. We set $\Delta_j - \Delta_{j-1} = 0.1$ ($j = 2, 3, \dots, w$) for a high-resolution mapping although this can be computationally burdensome due to the large number of values on the grid. For example, if the feasible range for Δ is $[-1500, 1500]$, the grid consists of 30,001 values. At one second per evaluation this requires about 8 hours and 20 minutes. Though computationally expensive, the high-resolution mapping of $L_{\text{prof}}(\Delta)$ is useful because it clearly identifies the likely (modal) values of Δ . In practice, we use multiple cores in parallel to reduce the computation time.

The profile likelihood evaluated on the grid can be used to approximate the posterior mean $E(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$ by

$$\hat{\Delta}_{\text{mean}} \equiv \frac{\sum_{j=1}^w \Delta_j \times L_{\text{prof}}(\Delta_j)}{\sum_{j=1}^w L_{\text{prof}}(\Delta_j)}, \quad (2.30)$$

and the posterior variance $\text{Var}(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$ by

$$\hat{V} \equiv \frac{\sum_{j=1}^w \Delta_j^2 \times L_{\text{prof}}(\Delta_j)}{\sum_{j=1}^w L_{\text{prof}}(\Delta_j)} - \left[\frac{\sum_{j=1}^w \Delta_j \times L_{\text{prof}}(\Delta_j)}{\sum_{j=1}^w L_{\text{prof}}(\Delta_j)} \right]^2. \quad (2.31)$$

Moreover, the posterior mode of Δ can be approximated by a value of Δ in the grid that maximizes the profile likelihood, which is a discrete approximation to the maximum likelihood estimator, $\hat{\Delta}_{\text{MLE}} \equiv \arg \max_{\Delta} L_{\text{prof}}(\Delta)$.

2.5 Time delay estimation strategy and numerical illustrations

The first step of our analysis is to plot $L_{\text{prof}}(\Delta)$ over the range of Δ to check for multimodality that may indicate multiple modes in the marginal posterior distribution of Δ . For some

quasars, the interesting range of Δ can be narrowed using the results of past analyses or information from other astrophysical probes. If prior information for Δ is unavailable, we explore the feasible range.

In our numerical studies, we find that when $L_{\text{prof}}(\Delta)$ is unimodal, the moment estimates of Δ based on $L_{\text{prof}}(\Delta)$, i.e., $\hat{\Delta}_{\text{mean}}$ in (2.30) and \hat{V} in (2.31), are almost identical to the posterior mean and variance obtained via MCMC. On the other hand, modes near the margins of the range of Δ may indicate microlensing, see Section 2.5.1. In this case, the order of polynomial regression must be increased. If there are multiple modes, but they are not near the margins of the range, each mode merits investigation.

As a cross-check, we run three MCMC chains near the major mode(s) identified by $L_{\text{prof}}(\Delta)$. The three starting values for each mode are {mode, mode \pm 20 days}. Each chain is run for 20,000 iterations and the first 10,000 iterations are discarded as burn-in. For all chains, we set the starting value of $\boldsymbol{\beta}$ to the estimated regression coefficients obtained by regressing $\mathbf{y}(\mathbf{t}) - \sum_j x(t_j)/n$ on a covariate matrix $\mathbf{W}_m(\mathbf{t} - \Delta^{(0)})$ whose j th row is $W_m^\top(t_j - \Delta^{(0)})$, where $\Delta^{(0)}$ is the initial value of Δ . The initial value of $\mathbf{X}(\mathbf{t}^\Delta)$ is the combined light curve, that is, $\{\mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t} - \Delta^{(0)}) - \mathbf{W}_m^\top(\mathbf{t} - \Delta^{(0)})\boldsymbol{\beta}^{(0)}\}$ sorted in time. The starting value of μ is set to the mean of $\mathbf{x}(\mathbf{t})$, that of σ^2 to 0.01^2 , and that of τ to 200. We set the initial standard deviations of the proposal distributions as $\psi = 10$ and $\phi = 3$.

We use simulated data of doubly- and quadruply-lensed quasars publicly available at the TDC website (<http://timedelaychallenge.org>) to illustrate our time delay estimation strategy when prior information for Δ is not available. We also analyze observed data of quasars *Q0957+561* and *J1029+2623* over the feasible range of Δ for illustrative purpose, though prior information is available to limit the range of Δ .

We report the CPU time in second using a server equipped with two 8-core Intel Xeon E5-2690 at 2.9 GHz and 64 GB of memory. We report the entire mapping time for $L_{\text{prof}}(\Delta)$.

2.5.1 A doubly-lensed quasar simulation

The simulated data for a doubly-lensed quasar are plotted in the first panel of Figure 2.7; the median cadence is 3 days, the cadence standard deviation is 1 day, and observations are made for 4 months in each of 5 years for 200 observations in total. The light curves suffer from microlensing which can be identified from their different long-term linear trends and similar short-term (intrinsic) variability.

To show the effect of microlensing on the time delay estimation, we fit both the curve-shifted model ($m = 0$) in (2.1) and the microlensing model with $m = 3$ in (3.15). We plot $\log(L_{\text{prof}}(\Delta))$ and $L_{\text{prof}}(\Delta)$ based on the curve-shifted model over the feasible range, $[t_1 - t_n, t_n - t_1] = [-1575.85, 1575.85]$, in the two panels of Figure 2.8. The profile likelihood exhibits large modes near the margins that overwhelm the profile likelihood near the true time delay (5.86 days denoted by the vertical red dashed line).

In the presence of microlensing, the curve-shifted model cannot identify the time delay because the latent curves are not shifted versions of each other. The modes of $L_{\text{prof}}(\Delta)$ near the margins of the range of Δ occur because a small overlap between the tips of two light curves may exhibit the only similar fluctuation patterns detectable by shifting one of the light curves. In Figure 2.9, for instance, we shift light curve B in the x -axis by the three

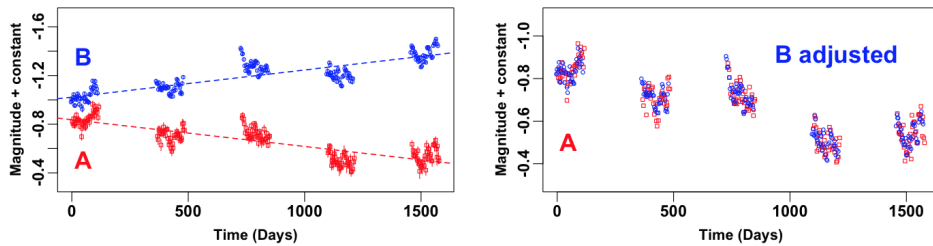


Figure 2.7: The first panel shows a TDC data set suffering from microlensing that results in light curves with different long-term trends. The dashed lines denote fitted linear regression lines. In the second panel, we shift light curve B by $\hat{\Delta}_{\text{MLE}}$ in the x -axis and subtract the estimated third-order polynomial regression from light curve B (using the values of β that maximize the profile likelihood at $\hat{\Delta}_{\text{MLE}}$). The microlensing model finds matches between the intrinsic fluctuations of the light curves after removing the relative microlensing trend from light curve B .

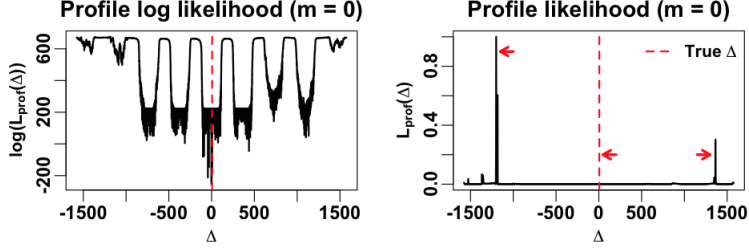


Figure 2.8: The profile log likelihood (left) and the profile likelihood (right) of Δ over its feasible range under the curve-shifted model ($m = 0$). We exponentiate $\log(L_{\text{prof}}(\Delta))$, setting the largest value of $L_{\text{prof}}(\Delta)$ to one. The vertical red dashed line indicates the true time delay. The profile likelihood near the true time delay (5.86 days) is overwhelmed by the modes near margins.

values of Δ indicated by three arrows in the second panel of Figure 2.8. In the first panel of Figure 2.9, the two light curves shifted by the true time delay do not match for any shift in magnitude. However, given the time delays at around $-1,200$ and $1,360$ days, the two light curves look well-connected as shown in the second and third panels. Thus, the profile likelihood near the true time delay is overwhelmed by the values of the profile likelihood near $-1,200$ and $1,360$ days.

To correct this effect, we fit the microlensing model with a third-order polynomial regression ($m = 3$). Both $\log(L_{\text{prof}}(\Delta))$ and $L_{\text{prof}}(\Delta)$ are plotted in Figure 2.10. One mode clearly

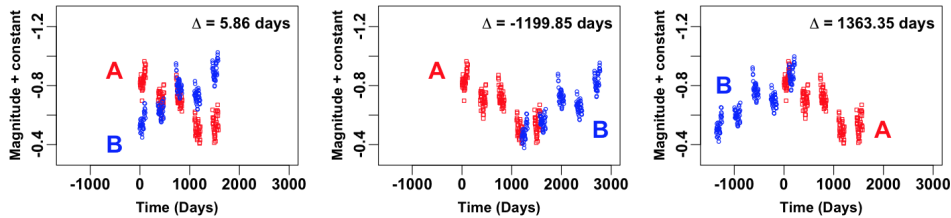


Figure 2.9: We shift light curve B (blue) by the true time delay (5.86 days) in the first panel, by $-1,199.85$ days in the second panel, and by 1363.35 days in the third panel. These three time delays correspond to three arrows in the second panel of Figure 2.8. The shift in magnitude used is the value of β_0 that maximizes the profile likelihood given each time delay. Without accounting for microlensing, the curve-shifted model fails because the light curves do not match even at the true time delay. The curve-shifted model may produce large modes near the margins because a small overlap between the tips of two light curves may have the only similar fluctuation patterns detectable by shifting one of the light curves as shown in the second and third panels.

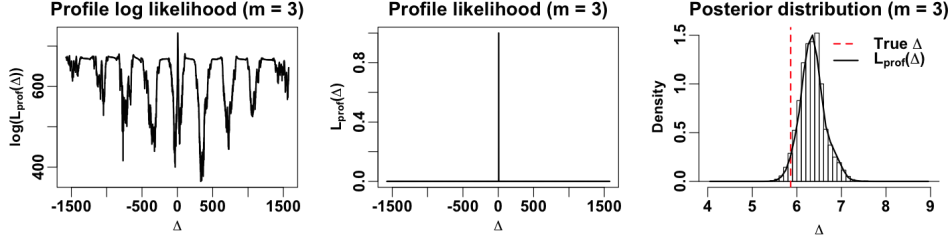


Figure 2.10: The profile log likelihood (first panel) and the profile likelihood (second panel) of Δ over its feasible range under the microlensing model ($m = 3$). The profile likelihood shows one mode near the true time delay (5.86 days). The third panel shows the marginal posterior distribution of Δ as a histogram of the MCMC samples with $L_{\text{prof}}(\Delta)$ superimposed. The vertical red dashed line indicates the true time delay.

dominates $L_{\text{prof}}(\Delta)$ and $\hat{\Delta}_{\text{mean}} = 6.36$ days. Using a uniform prior for Δ over its feasible range and $\sigma^2 \sim \text{IG}(1, 2/10^7)$, we initialize three MCMC chains near 6.36 days. We check the convergence with the Gelman-Rubin diagnostic statistic (GRD, Gelman and Rubin, 1992) which equals 1.0007 for Δ . It took 14,457 seconds to map $L_{\text{prof}}(\Delta)$ and 186 seconds on average for each MCMC chain. The profile likelihood and marginal posterior near the dominant mode are almost identical and are consistent with the true value of Δ as shown in the third panel of Figure 2.10.

In the second panel of Figure 2.7, we shift light curve B by $\hat{\Delta}_{\text{MLE}}$ (x -axis) and subtract the estimated polynomial regression (using the values of β that maximize $L_{\text{prof}}(\Delta)$ evaluated at $\hat{\Delta}_{\text{MLE}}$). The microlensing model finds matches between the intrinsic fluctuations of the light curves after removing the relative microlensing trend from light curve B .

We summarize the Bayesian and profile likelihood estimates for Δ in Table 2.1. The true

Table 2.1: Estimates of Δ ; the profile likelihood estimates, $\hat{\Delta}_{\text{mean}}$ and $\hat{V}^{0.5}$ are given in the $E(\Delta|\mathbf{x}(t), \mathbf{y}(t))$ and $SD \equiv SD(\Delta|\mathbf{x}(t), \mathbf{y}(t))$ columns, where $Error \equiv |\Delta_{\text{true}} - E(\Delta|\mathbf{x}(t), \mathbf{y}(t))|$ with Δ_{true} indicating the true time delay (5.86 days), and $\chi \equiv Error/SD(\Delta|\mathbf{x}(t), \mathbf{y}(t))$.

| Method | $E(\Delta D_{\text{obs}})$ | $\hat{\Delta}_{\text{MLE}}$ | SD | Δ_{true} | Error | χ |
|--------------------|----------------------------|-----------------------------|------|------------------------|-------|--------|
| Bayesian | 6.34 | | 0.28 | 5.86 | 0.48 | 1.71 |
| Profile likelihood | 6.36 | 6.35 | 0.28 | 5.86 | 0.50 | 1.76 |

delay is within two posterior standard deviation ($\chi \leq 1.76$) of the posterior mean; similar accuracy is obtained with the profile likelihood approximation.

2.5.2 A quadruply-lensed quasar simulation

The simulated data for a quadruply-lensed quasar are plotted in Figure 2.11 and are composed of four light curves, A, B, C , and D ; the median cadence is 6 days, the cadence standard deviation is 1 day, and observations are made for 4 months in each of 10 years with 200 observations in total. The feasible range for each Δ is $[-3391.62, 3391.62]$.

With quadruply lensed data there are six time delay parameters, one for each pair of light curves. Any three of these parameters determine the others and we focus on Δ_{AB} , Δ_{AC} , and Δ_{AD} , where the subscripts index the two light curves being compared. This pair-wise approach proceeds by applying the method developed for doubly-lensed data in Section 2.5.1 to the pair of light curves corresponding to each of Δ_{AB} , Δ_{AC} , and Δ_{AD} in turn (Fassnacht et al., 1999).

A coherent model would consider all four time series data in one model simultaneously (Hojjati et al., 2013; Tewes et al., 2013); four time series are generated from one underlying true process and the three distinct time delays may have a posteriori correlations. By focusing on pairwise comparisons of the four time series, we do not account for the correlations between the time delays. Extending our model to simultaneously consider all of the data is a topic for future research.

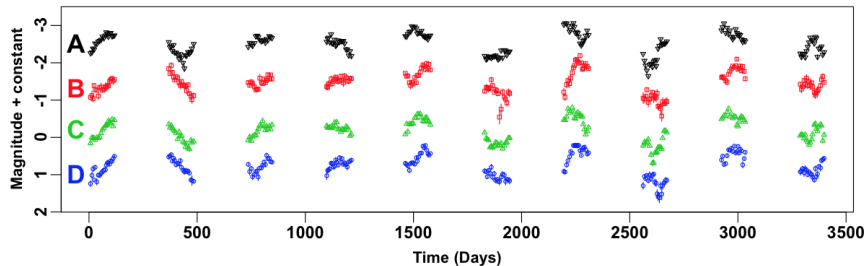


Figure 2.11: Simulated quadruply-lensed quasar data used in the TDC.

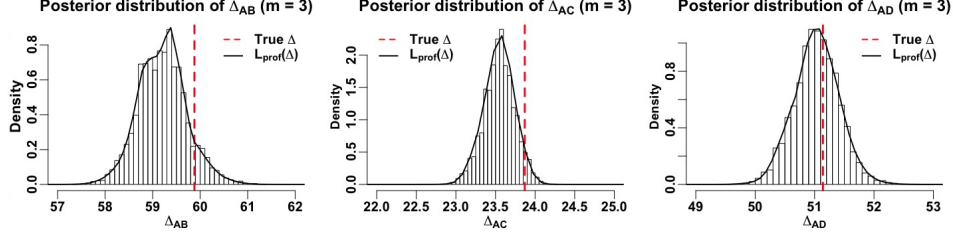


Figure 2.12: The marginal posterior distributions of Δ_{AB} (first panel), Δ_{AC} (second panel), and Δ_{AD} (third panel) with $L_{\text{prof}}(\Delta)$ superimposed. Vertical red dashed lines indicate blinded true time delays.

We follow our default strategy to analyze these simulated data, using the microlensing model ($m = 3$). After confirming a single dominating mode in the profile likelihood for each time delay parameter, we initiate three MCMC chains near this mode; the GRD is 1.0000 for Δ_{AB} , 1.0010 for Δ_{AC} , and 1.0001 for Δ_{AD} . The posterior distributions of Δ_{AB} , Δ_{AC} , and Δ_{AD} appear in Figure 2.12 with $L_{\text{prof}}(\Delta)$ superimposed. The profile likelihood is almost identical to the posterior distribution of each parameter and both estimate the true time delays well. The average CPU time taken to map $L_{\text{prof}}(\Delta)$ is about 73,000 seconds (averaging over the three time delays). The average CPU time taken for each MCMC chain is about 200 seconds (averaging over nine chains; three chains for each time delay). Our estimation results are summarized in Table 2.2. The Bayesian estimates and profile likelihood approximations are

Table 2.2: Estimates of Δ_{AB} , Δ_{AC} , and Δ_{AD} ; the profile likelihood estimates, $\hat{\Delta}_{\text{mean}}$ and $\hat{V}^{0.5}$ are given in the $E(\Delta|\mathbf{x}(t), \mathbf{y}(t))$ and $SD \equiv SD(\Delta|\mathbf{x}(t), \mathbf{y}(t))$ columns, where $Error \equiv |\Delta_{\text{true}} - E(\Delta|\mathbf{x}(t), \mathbf{y}(t))|$ with Δ_{true} indicating the true time delay, i.e., $\Delta_{AB} = 59.88$, $\Delta_{AC} = 23.87$ and $\Delta_{AD} = 51.14$, and $\chi \equiv Error/SD(\Delta|\mathbf{x}(t), \mathbf{y}(t))$.

| | Method | $E(\Delta \mathbf{x}(t), \mathbf{y}(t))$ | $\hat{\Delta}_{\text{MLE}}$ | SD | Δ_{true} | Error | χ |
|---------------|--------------------|--|-----------------------------|------|------------------------|-------|--------|
| Δ_{AB} | Bayesian | 59.21 | | 0.52 | 59.88 | 0.67 | 1.29 |
| | Profile likelihood | 59.21 | 59.38 | 0.51 | 59.88 | 0.67 | 1.31 |
| Δ_{AC} | Bayesian | 23.55 | | 0.20 | 23.87 | 0.32 | 1.60 |
| | Profile likelihood | 23.54 | 23.58 | 0.19 | 23.87 | 0.33 | 1.74 |
| Δ_{AD} | Bayesian | 51.03 | | 0.39 | 51.14 | 0.11 | 0.28 |
| | Profile likelihood | 51.03 | 51.08 | 0.38 | 51.14 | 0.11 | 0.29 |

quite similar and both produce estimates within two standard deviations of the truth (all $\chi \leq 1.74$).

2.5.3 Quasar *Q0957+561*

The first known gravitationally (doubly) lensed quasar *Q0957+561* was discovered by Walsh et al. (1979) who suggested that a strong gravitational lensing may have formed the two images. Here we analyze the most recent observation of this quasar. This observation was made by the United States Naval Observatory in 2008–2011 (Hainline et al., 2012). The data were observed on 57 nights and are plotted in the first panel of Figure 2.13. The feasible range for Δ is $[-1178.939, 1178.939]$.

Inspection of $L_{\text{prof}}(\Delta)$ revealed one dominant mode. Using a uniform prior distribution of Δ over its range and $\sigma^2 \sim \text{IG}(1, 2/10^7)$, we ran three MCMC chains near this mode; the GRD for Δ is 1.0001. The second panel of Figure 2.13 shows the marginal posterior distribution of Δ with $L_{\text{prof}}(\Delta)$ superimposed. It took 1,243 seconds to map $L_{\text{prof}}(\Delta)$ and 71 seconds on average for each MCMC chain.

In this case the dominant mode of Δ has more complex structure as detected by both $L_{\text{prof}}(\Delta)$ and the MCMC sample, see the second panel of Figure 2.13. In the third panel of Figure 2.13, we shift light curve *B* by $\hat{\Delta}_{\text{MLE}}$ (x -axis) and subtract the estimated third-order

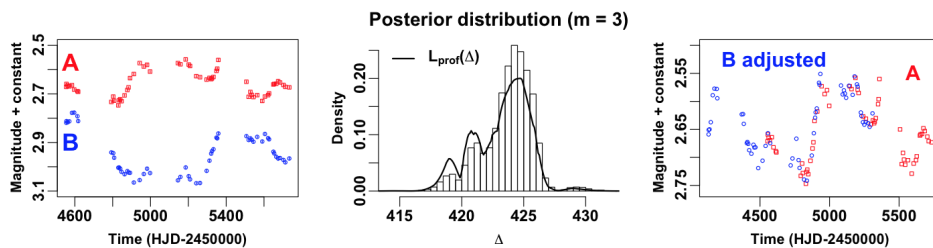


Figure 2.13: Observations of Quasar *Q0957+561* from Hainline et al. (2012) are plotted in the first panel. The second panel exhibits the marginal posterior distribution of Δ with $L_{\text{prof}}(\Delta)$ superimposed. We shift light curve *B* by $\hat{\Delta}_{\text{MLE}}$ (x -axis) and subtract the estimated third-order polynomial regression (obtained by the profile likelihood at $\hat{\Delta}_{\text{MLE}}$) in the third panel. HJD indicates the Heliocentric Julian date.

Table 2.3: Historical time delay estimates and standard errors (SE) for Q0957+561 (*r*-band). Our work provides the posterior mean and standard deviation of Δ and profile likelihood approximations to them. Serra-Ricart et al. (1999), Oscoz et al. (1997, 2001) and Pelt et al. (1996) used a bootstrapping method to calculate the SE.

| Researchers | Method | Estimate | SE |
|----------------------------|---|----------|------|
| Pelt et al. (1996) | Dispersion method | 423 | 6 |
| Oscoz et al. (1997) | Discrete cross-correlation & Dispersion | 424 | 3 |
| Serra-Ricart et al. (1999) | Least square optimization via auto-& cross-correlation | 425 | 4 |
| Oscoz et al. (2001) | Discrete correlation function (DCF) | 426 | 5 |
| | Discrete auto-& cross-correlation | 423 | 2 |
| | Z-transformed DCF | 420 | 8 |
| | χ^2 -minimization | 422 | 3 |
| This work | Bayesian | 423.73 | 2.02 |
| | Profile likelihood | 423.21 | 2.81 |

polynomial regression. The fitted microlensing model matches the intrinsic fluctuations of the two light curves.

Estimates based on different observations of Q0957+561 appear in Table 2.3. Although the observations span different time periods, the resulting estimates, including ours, are broadly consistent. Though the posterior mean and standard deviation may be difficult to interpret with a multimodal posterior distribution, we include them for comparison.

2.5.4 Quasar J1029+2623

Inada et al. (2006) discovered the gravitationally lensed quasar *J1029+2623* whose time delay is the second largest yet observed. Though *J1029+2623* has three images (A, B, and C), Fohlmeister et al. (2013) merged B and C because C is indistinguishable from B due to its faintness and proximity. They published its data (A, B+C) with 279 epochs monitored at the Fred Lawrence Whipple observatory from January 2007 to June 2012. The first panel

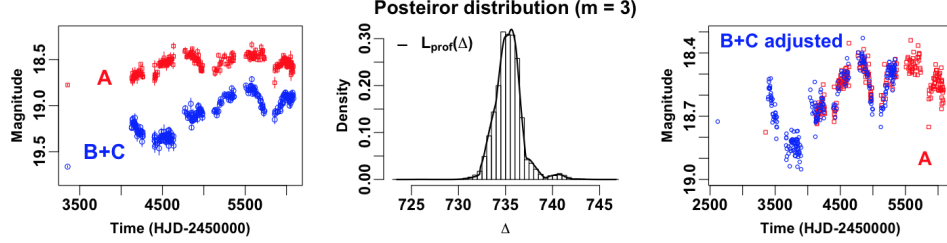


Figure 2.14: We plot the observations of Quasar J1029+2623 from Fohlmeister et al. (2013) in the first panel. The second panel exhibits the marginal posterior distribution of Δ with $L_{\text{prof}}(\Delta)$ superimposed. We shift light curve $B + C$ by $\hat{\Delta}_{\text{MLE}}$ (x -axis) and subtract the estimated third-order polynomial regression (obtained by the profile likelihood at $\hat{\Delta}_{\text{MLE}}$) in the last panel. HJD indicates the Heliocentric Julian date.

in Figure 2.14 shows these data. The feasible range of Δ is $[-2729.759, 2729.759]$.

We confirmed a dominant mode via $L_{\text{prof}}(\Delta)$ and using a uniform prior distribution of Δ over its range and $\sigma^2 \sim \text{IG}(1, 2/10^7)$, we initiated three MCMC chains near this mode; the GRD for Δ is 1.0009. We display the marginal posterior distribution of Δ in the second panel of Figure 2.14 with $L_{\text{prof}}(\Delta)$ superimposed. It took 33,683 seconds to map $L_{\text{prof}}(\Delta)$ and 311 seconds on average for each MCMC chain. The posterior distribution and the profile likelihood are almost identical. In the third panel, we shift light curve B by $\hat{\Delta}_{\text{MLE}}$ (x -axis) and subtract the estimated third-order polynomial regression. Again, the fitted microlensing model finds a good match of the two light curves.

Table 2.4: Historical time delay estimates and 90% confidence intervals for J1029+2623. Our work provides the posterior mean and 90% posterior interval of Δ and profile likelihood approximations to them. Fohlmeister et al. (2013) did not specify how they produced the sampling distribution of Δ . Kumar et al. (2014) used a parametric bootstrapping method.

| Researchers | Method | Estimate | 90% Interval |
|-----------------------------------|-----------------------------------|----------|------------------|
| Fohlmeister et al. (2013) | χ^2 -minimization (AIC, BIC) | 744 | (734, 754) |
| Kumar, Stalin, and Prabhhu (2014) | Difference-smoothing | 743.5 | (734.6, 752.4) |
| This work | Bayesian | 735.31 | (733.08, 737.71) |
| | Profile likelihood | 733.11 | (732.94, 738.44) |

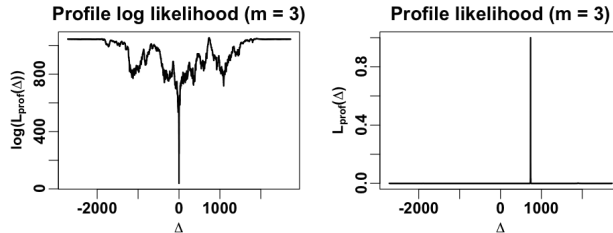


Figure 2.15: The profile log likelihood (left) and the profile likelihood (right) of Δ over its feasible range under the microlensing model ($m = 3$). Although the profile likelihood shows a dominant mode near 735 days, there are small modes near -2,000 and 1,800 days.

We compare our estimates in Table 2.4 with historical estimates that are based on the same data. The Bayesian method uses 5% and 95% quantiles of the posterior samples of Δ as the 90% posterior interval. To obtain the 90% interval estimate for Δ via the profile likelihood, we draw 50,000 samples of Δ using the empirical CDF of the (normalized) profile likelihood and choose their 5% and 95% quantiles.

The shape of the posterior distribution of Δ is almost identical to that of the profile likelihood in the second panel of Figure 2.14. However, the posterior mean of Δ is larger than the profile approximation, $\hat{\Delta}_{\text{mean}}$, by about two days. This is because there are two small modes near -2,000 and 1,800 days, see the second panel of Figure 2.15. Because the mode near 735 days overwhelmed the other modes, it is reasonable to focus on the mode near 735 days as the Bayesian result does. Overall, our point estimates are smaller than the historical estimates by about ten days, though our 90% posterior intervals overlap with the other historical 90% confidence intervals in Table 2.4.

2.6 Conclusion

Accurately estimating time delays among gravitationally lensed quasar images is a key to making fundamental measurements of the current expansion rate of the Universe and dark energy (Refsdal, 1964; Linder, 2011). The Large Synoptic Survey Telescope (LSST Science Collaboration, 2009) will produce extensive time series data on thousands of multiply lensed

quasars starting in 2022. Anticipating this era of astronomical Big Data, we have improved the fully Bayesian model of Harva and Raychaudhury (2006) by leveraging recent advances in astrophysical and statistical modeling. We have added an Orstein-Uhlenbeck process to model the fluctuations in quasar light curves, a polynomial regression to account for microlensing, and a profile-likelihood-guided Bayesian strategy.

There are several opportunities to build upon our work in future research. It is desirable to implement more sophisticated methods of model selection such as information criteria to choose the complexity of the microlensing trend. Though astrophysicists have used a cubic polynomial trend for microlensing models for some quasars so far, it would be better to have a fast and principled mechanism to determine the order given any data of gravitationally lensed quasars. Another avenue for further improvement is to constrain the range of the time delay by incorporating additional astrophysical information such as spatial positions of the images relative to the lensing galaxy, and an astrophysical model for the mass distribution of the lens. For quadruply-lensed quasar systems, constructing a Bayesian model to simultaneously analyze the four light curves, would allow us to coherently estimate the relative time delays without loss of information. Further improvements to the computational efficiency of our profile likelihood and MCMC strategies will enhance their effectiveness for analyzing the extensive time series datasets expected in the era of the LSST.

Chapter 3

A Repelling-Attracting Metropolis Algorithm for Multimodality

3.1 Introduction

Multimodal distributions are common in statistical applications. However, a Metropolis algorithm, one of the most widely used Markov chain Monte Carlo (MCMC) methods, tends to produce a Markov chain being stuck at one of the local modes for a long time when a target distribution has several modes. A popular MCMC strategy for dealing with multimodality is tempering. Tempering melts down the modes of a target density to create a flatter surface and hence improved mixing. There are many temperature-based methods such as parallel tempering (Geyer, 1991), simulated tempering (Geyer and Thompson, 1995), tempered transitions (Neal, 1996), and equi-energy sampler (Kou et al., 2006). Though powerful, these methods typically require extensive tuning and tend to be computationally expensive.

We use a Metropolis algorithm as a building block to construct an alternative, easy-to-implement and temperature-free multimodal sampler called a repelling-attracting Metropolis (RAM) algorithm. This algorithm enables a Markov chain to jump between modes more frequently than a Metropolis algorithm with less tuning than any tempering methods. The

RAM algorithm generates a proposal via forced downhill and forced uphill Metropolis transitions. The term *forced* emphasizes that neither Metropolis transition is allowed to stay at its current state because we repeatedly make proposals until one is accepted. The forced downhill Metropolis transition uses a reciprocal ratio of the target densities in its acceptance probability. This encourages the intermediate proposal to prefer downward moves since a lower density state has a higher chance of acceptance, hence local modes become *repelling*. The subsequent forced uphill Metropolis transition generates a final proposal with a standard Metropolis ratio that makes local modes *attracting*. Together the downhill and uphill transitions form a proposal for a Metropolis-Hastings (M-H) sample; a final accept-reject step preserves the target stationary distribution. The final proposal has a higher chance to be in a mode other than the one of the current state, as shown in Figure 3.1, and it is then accepted or rejected in the usual way.

The scale of the proposal distributions, either a scalar or a matrix depending on dimensionality, iterated within the downhill and uphill Metropolis transitions is the only tuning parameter of this algorithm if the proposal distributions are Gaussian. It is easy to fine-tune the proposal scale if the information about the locations of modes are known, but the RAM algorithm does not necessarily need such information. As with other M-H samplers, the normalizing constant of the target density need not be known. Consequently, it is always possible to replace a Metropolis algorithm with the RAM algorithm to explore a multimodal

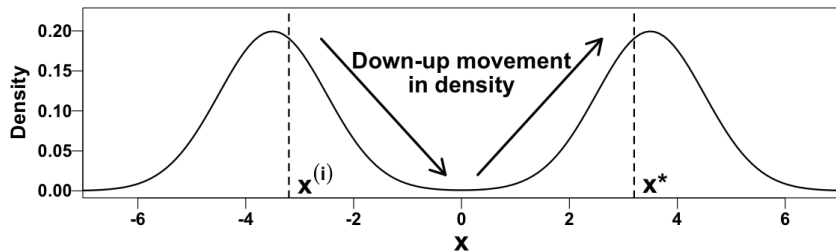


Figure 3.1: A repelling-attracting Metropolis algorithm is a Metropolis-Hastings algorithm that generates a proposal x^* given the current state $x^{(i)}$ by making a down-up movement in density, i.e., repelling-attracting to local modes, via forced downhill and uphill Metropolis transitions.

distribution better because running the RAM algorithm does not require more information than what is needed to run a Metropolis algorithm.

Although we can draw a sample using the down-up proposal rule, the acceptance probability of the final proposal contains a ratio of intractable integrals. We solve this problem by introducing an auxiliary variable, using the idea of Møller et al. (2006). This auxiliary variable approach marginally preserves the target density and requires another forced downhill Metropolis transition for the auxiliary variable. Thus, the RAM algorithm generates a proposal via three forced Metropolis transitions but accepts the proposal with an easy-to-compute acceptance probability.

Using several numerical examples, we show a benefit of simply replacing a Metropolis algorithm with the RAM algorithm in exploring a multimodal distribution. We also compare the performance of the RAM algorithm to that of commonly-used tempering methods, e.g., parallel tempering, equi-energy sampler, and tempered transitions. In the first example, we look into a high-dimensional behavior of the RAM algorithm compared to that of a Metropolis algorithm in exploring a mixture of three d -dimensional Gaussian distributions for $d \in \{3, 10, 20\}$. The target distribution in the second example is a mixture of 20 bivariate Gaussian distributions with either equal-variance and equally-weighted modes or unequal-variance and unequally-weighted modes (Kou et al., 2006). In this example, we show that the mean squared error of moment estimates from the RAM algorithm is better than that of both parallel tempering and equi-energy sampler. The last example is from our applied work in astrophysics, which has motivated this research. Here, we show that the RAM algorithm is better than tempered transitions in exploring a grossly multimodal distribution whose modes are more than 600 standard deviations distant from each other.

3.2 A repelling-attracting Metropolis algorithm

We briefly review the Metropolis-Hastings (M-H) algorithm. A transition kernel on \mathbf{R}^d , denoted by $P(B | x)$, is the conditional probability distribution of transition from $x \in \mathbf{R}^d$ to a point in a Borel set B in \mathbf{R}^d ; $P(\mathbf{R}^d | x) = 1$ and $P(\{x\} | x)$ need not be zero (Chib and Greenberg, 1995). A proposal density given the current state $x^{(i)}$ is a conditional density that generates a proposal x^* . We denote this proposal density by $q(x^* | x^{(i)})$ which must satisfy $\int q(x^* | x^{(i)})dx^* = 1$. With a target density denoted by π , either normalized or unnormalized, a transition kernel of the M-H algorithm is defined as

$$P(dx^* | x^{(i)}) = q(x^* | x^{(i)})\alpha(x^* | x^{(i)})dx^* + \delta_{x^{(i)}}(dx^*)\{1 - A(x^{(i)})\}, \quad (3.1)$$

where $\alpha(x^* | x^{(i)})$ is the probability of accepting the proposal x^* as $x^{(i+1)}$, i.e.,

$$\alpha(x^* | x^{(i)}) = \min \left\{ 1, \frac{\pi(x^*)q(x^{(i)} | x^*)}{\pi(x^{(i)})q(x^* | x^{(i)})} \right\};$$

$1 - A(x^{(i)})$ is the probability of staying at $x^{(i)}$ and thus $A(x^{(i)})$ is that of moving from $x^{(i)}$,

$$A(x^{(i)}) = \int q(x^* | x^{(i)})\alpha(x^* | x^{(i)})dx^*;$$

and the Dirac measure $\delta_{x^{(i)}}(dx^*)$ is one if $x^{(i)} \in dx^*$ and zero otherwise. If the proposal density is symmetric, satisfying $q(x^* | x^{(i)}) = q(x^{(i)} | x^*)$, then the M-H algorithm reduces to a Metropolis algorithm, whose acceptance probability is

$$\alpha(x^* | x^{(i)}) = \min \left\{ 1, \frac{\pi(x^*)}{\pi(x^{(i)})} \right\}. \quad (3.2)$$

The RAM algorithm is essentially a M-H algorithm with a proposal density that is designed to boost the down-up movement via two forced Metropolis transitions. The forced Metropolis algorithm is the same as a standard Metropolis algorithm except that the forced algorithm repeatedly makes proposals until one is accepted. Without a forced transition, the final proposal x^* could be the same as the current state $x^{(i)}$ after consecutive rejections in both the downhill and uphill Metropolis transitions, which is wasteful because the

final proposal $x^* = x^{(i)}$ is accepted for certain. Also, if the forced transitions were not included, the final proposal would be generated via only one of the two Metropolis transitions if the other were rejected. This would not be helpful for our purposes because it would not induce a down-up movement. Moreover, a transition kernel of the forced Metropolis algorithm is mathematically easier to handle than that of a Metropolis algorithm. This is because the transition kernel of the forced Metropolis algorithm is no longer a mixture of two distributions; the second mixture component for staying at the current state, e.g., $\delta_{x^{(i)}}(dx^*)\{1 - A(x^{(i)})\}$ in (3.1), disappears.

The forced downhill Metropolis transition generates an intermediate proposal x' from the current state $x^{(i)}$ using the reciprocal ratio of the target densities in its acceptance probability,

$$\alpha_\epsilon^D(x' | x^{(i)}) = \min \left\{ 1, \frac{\pi(x^{(i)}) + \epsilon}{\pi(x') + \epsilon} \right\}, \quad (3.3)$$

where the superscript, D, indicates that the ratio has been flipped for a downward move; the appearance of ϵ in (3.3) is discussed below. The reciprocal density ratio in (3.3) makes local modes repelling rather than attracting: If the density of x' is smaller than that of $x^{(i)}$, x' is accepted with probability one. The forced uphill Metropolis transition restores the attractiveness of local modes as with the original Metropolis ratio which prefers upward movement in density. The forced uphill Metropolis transition generates the final proposal x^* given x' , whose acceptance probability is

$$\alpha_\epsilon^U(x^* | x') = \min \left\{ 1, \frac{\pi(x^*) + \epsilon}{\pi(x') + \epsilon} \right\}, \quad (3.4)$$

where the superscript, U, indicates that the acceptance probability prefers an upward movement. The acceptance probability in (3.4) is the same as in (3.2) except that ϵ is added to the numerator and denominator. This is done for numerical stability; both $\pi(x')$ and $\pi(x^*)$ can be nearly zero when both x' and x^* are in a flat valley between two distant modes. In this case, adding ϵ prevents a ratio of zeros in the acceptance probability. However, ϵ may affect the convergence rate of the sampler because a large value of ϵ that dominates π results

in x^* almost always being accepted, regardless of whether x^* is an uphill move or not. To minimize its impact on the acceptance probability in (3.4), ϵ must be small and our default choice is $\epsilon = 10^{-308}$, a constant that R treats as positive (R Development Core Team, 2015); R treats $1/10^{309}$ as zero. For a symmetry, we also use ϵ in the same way in the acceptance probability of the downhill transition in (3.3), which leads to a symmetry of the acceptance probability up to D and U, i.e.,

$$\alpha_\epsilon^D(x' | x^{(i)}) = \min \left\{ 1, \frac{\pi(x^{(i)}) + \epsilon}{\pi(x') + \epsilon} \right\} = \alpha_\epsilon^U(x^{(i)} | x'), \quad (3.5)$$

$$\alpha_\epsilon^U(x^* | x') = \min \left\{ 1, \frac{\pi(x^*) + \epsilon}{\pi(x') + \epsilon} \right\} = \alpha_\epsilon^D(x' | x^*). \quad (3.6)$$

Thus, the RAM algorithm is a M-H algorithm with a down-up proposal density

$$q^{\text{DU}}(x^* | x^{(i)}) = \int q^D(x' | x^{(i)}) q^U(x^* | x') dx', \quad (3.7)$$

where q^D and q^U are the forced downhill and uphill transition kernel densities, respectively.

Specifically, the forced downhill kernel density is

$$q^D(x' | x^{(i)}) = \frac{q(x' | x^{(i)}) \alpha_\epsilon^D(x' | x^{(i)})}{A^D(x^{(i)})}, \quad (3.8)$$

$$A^D(x^{(i)}) = \int q(x' | x^{(i)}) \alpha_\epsilon^D(x' | x^{(i)}) dx',$$

where $A^D(x^{(i)})$ is the probability of accepting any single proposal from $q(x' | x^{(i)})$. Similarly, the forced uphill kernel density is

$$q^U(x^* | x') = \frac{q(x^* | x') \alpha_\epsilon^U(x^* | x')}{A^U(x')}, \quad (3.9)$$

$$A^U(x') = \int q(x^* | x') \alpha_\epsilon^U(x^* | x') dx^*.$$

Consequently, the down-up proposal density in (3.7) satisfies $\int q^{\text{DU}}(x^* | x^{(i)}) dx^* = 1$.

The conditional density q in (3.8) and (3.9) may be any symmetric density to simplify RAM's final acceptance probability, and it must have a positive probability of reaching out all possible states for irreducibility. For instance, we can set $q(a | b) = N_d(a | b, \Sigma)$, a d -dimensional Gaussian density of a with mean b and variance-covariance matrix Σ , where Σ

is the only tuning parameter that can be used to improve the mixing of the RAM algorithm. A random-walk proposal can be made by setting Σ to a diagonal matrix and its simplest form is $\Sigma = \sigma^2 I_d$, where I_d is a $d \times d$ identity matrix and σ is a proposal scale. Since a random-walk proposal performs poorly as dimensionality increases, we improve RAM's high-dimensional behavior by setting Σ as follows:

$$\Sigma = \begin{cases} S_0 & \text{during a burn-in period,} \\ S & \text{after a burn-in period,} \end{cases} \quad (3.10)$$

where S_0 is an initial $d \times d$ variance-covariance matrix set by users, e.g., $S_0 = \sigma^2 I_d$, and S is a $d \times d$ sample variance-covariance matrix calculated from the samples drawn during the burn-in period.

Given the proposal, the M-H acceptance ratio is

$$\alpha^{\text{DU}}(x^* | x^{(i)}) = \min \left\{ 1, \frac{\pi(x^*)q^{\text{DU}}(x^{(i)} | x^*)}{\pi(x^{(i)})q^{\text{DU}}(x^* | x^{(i)})} \right\} = \min \left\{ 1, \frac{\pi(x^*)A^{\text{D}}(x^{(i)})}{\pi(x^{(i)})A^{\text{D}}(x^*)} \right\}. \quad (3.11)$$

The second equation in (3.11) holds because the symmetry of q , of $\alpha_\epsilon^{\text{D}}(x' | x^{(i)}) = \alpha_\epsilon^{\text{U}}(x^{(i)} | x')$ in (3.5), and of $\alpha_\epsilon^{\text{U}}(x^* | x') = \alpha_\epsilon^{\text{D}}(x' | x^*)$ in (3.6) implies the following relationship:

$$\begin{aligned} q^{\text{DU}}(x^* | x^{(i)})A^{\text{D}}(x^{(i)}) &= \int q(x' | x^{(i)})\alpha_\epsilon^{\text{D}}(x' | x^{(i)})\frac{q(x^* | x')\alpha_\epsilon^{\text{U}}(x^* | x')}{A^{\text{U}}(x')}dx' \\ &= \int q(x^{(i)} | x')\alpha_\epsilon^{\text{U}}(x^{(i)} | x')\frac{q(x' | x^*)\alpha_\epsilon^{\text{D}}(x' | x^*)}{A^{\text{U}}(x')}dx' \\ &= q^{\text{DU}}(x^{(i)} | x^*)A^{\text{D}}(x^*). \end{aligned}$$

Unfortunately, the acceptance probability in (3.11) is difficult to compute due to its ratio of intractable normalizing constants of the forced downhill kernel densities, $A^{\text{D}}(x^{(i)})/A^{\text{D}}(x^*)$. Møller et al. (2006) use an auxiliary variable approach to cancel out a ratio of intractable normalizing constants of a target density. We follow this approach, but our case arises from the intractable down-up proposal density, q^{DU} , which is a conditional density unlike the target density, π . We introduce an auxiliary variable in such a way that the marginal target density for x remains π exactly. This auxiliary variable results in a term that cancels with the intractable ratio.

Specifically, let $z \in \mathbf{R}^d$ be an auxiliary variable that shares the same space with x , via a conditional density $\pi^C(z | x)$ to be specified. We denote a joint proposal density that proposes (z^*, x^*) given the current states $(z^{(i)}, x^{(i)})$ by $q^J(z^*, x^* | z^{(i)}, x^{(i)})$ and assume that it factors and can be simplified as

$$q^J(z^*, x^* | z^{(i)}, x^{(i)}) = q_1(x^* | z^{(i)}, x^{(i)})q_2(z^* | x^*, z^{(i)}, x^{(i)}) = q_1(x^* | x^{(i)})q_2(z^* | x^*)$$

so that the M-H acceptance probability for the joint proposal is

$$\begin{aligned} \alpha^J(z^*, x^* | z^{(i)}, x^{(i)}) &= \min \left\{ 1, \frac{\pi(x^*)\pi^C(z^* | x^*)q^J(z^{(i)}, x^{(i)} | z^*, x^*)}{\pi(x^{(i)})\pi^C(z^{(i)} | x^{(i)})q^J(z^*, x^* | z^{(i)}, x^{(i)})} \right\} \\ &= \min \left\{ 1, \frac{\pi(x^*)\pi^C(z^* | x^*)q_1(x^{(i)} | x^*)q_2(z^{(i)} | x^{(i)})}{\pi(x^{(i)})\pi^C(z^{(i)} | x^{(i)})q_1(x^* | x^{(i)})q_2(z^* | x^*)} \right\}. \end{aligned} \quad (3.12)$$

This acceptance probability may resemble that of a pseudo-marginal approach (Beaumont, 2003; Andrieu and Roberts, 2009) that focuses on an unbiased estimator for an intractable target density. However, our auxiliary variable approach is different from a pseudo-marginal approach because our problem arises from an intractable proposal density and a tractable target density. From the pseudo-marginal perspective, the factorization of the joint target density in (3.12), i.e., $\pi(x, z) = \pi(x)\pi^C(z | x)$, is not allowed because it leaves the intractable target density $\pi(x)$ in the acceptance probability again.

Suppose it is possible to draw a sample from q_1 but difficult to evaluate q_1 . We can find a function f such that $q_1(x^{(i)} | x^*)/q_1(x^* | x^{(i)}) = f(x^{(i)})/f(x^*)$ because the ratio of two (compatible) conditional densities is the ratio of two marginal densities. The function f may or may not be computable and can be a normalizing constant of q_1 but not necessarily. If we can find a function q_2 whose normalizing constant is proportional to f , then the joint acceptance probability in (3.12) becomes free of the intractable quantities.

For the RAM algorithm, we set $q_1(x^* | x^{(i)}) = q^{\text{DU}}(x^* | x^{(i)})$ to propose a down-up movement from $x^{(i)}$ to x^* , where q^{DU} is specified in (3.7). In this case, $f(x^{(i)}) = A^{\text{D}}(x^{(i)})$ which is the normalizing constant of the forced downhill kernel density q^{D} in (3.8). To eliminate this intractable normalizing constant, we choose $q_2(z^* | x^*) = q^{\text{D}}(z^* | x^*)$. Møller

et al. (2006) suggest choosing π^C similar to q_2 and thus we assume $\pi^C(z^* | x^*)$ equals $q(z^* | x^*)$. With these choices, the acceptance probability in (3.12) reduces to

$$\begin{aligned}
\alpha^J(z^*, x^* | z^{(i)}, x^{(i)}) &= \min \left\{ 1, \frac{\pi(x^*)q(z^* | x^*)q^{\text{DU}}(x^{(i)} | x^*)q^{\text{D}}(z^{(i)} | x^{(i)})}{\pi(x^{(i)})q(z^{(i)} | x^{(i)})q^{\text{DU}}(x^* | x^{(i)})q^{\text{D}}(z^* | x^*)} \right\} \\
&= \min \left\{ 1, \frac{\pi(x^*)q(z^* | x^*)A^{\text{D}}(x^{(i)})q(z^{(i)} | x^{(i)})\alpha_\epsilon^{\text{D}}(z^{(i)} | x^{(i)})/A^{\text{D}}(x^{(i)})}{\pi(x^{(i)})q(z^{(i)} | x^{(i)})A^{\text{D}}(x^*)q(z^* | x^*)\alpha_\epsilon^{\text{D}}(z^* | x^*)/A^{\text{D}}(x^*)} \right\} \\
&= \min \left\{ 1, \frac{\pi(x^*)\alpha_\epsilon^{\text{D}}(z^{(i)} | x^{(i)})}{\pi(x^{(i)})\alpha_\epsilon^{\text{D}}(z^* | x^*)} \right\} = \min \left\{ 1, \frac{\pi(x^*) \min\{1, \frac{\pi(x^{(i)})+\epsilon}{\pi(z^{(i)})+\epsilon}\}}{\pi(x^{(i)}) \min\{1, \frac{\pi(x^*)+\epsilon}{\pi(z^*)+\epsilon}\}} \right\}. \quad (3.13)
\end{aligned}$$

In (3.13), $\pi(z^{(i)})$ is likely to be smaller than $\pi(x^{(i)})$ because $z^{(i)}$ is generated by the forced downhill transition. Similarly, $\pi(z^*)$ is likely to be smaller than $\pi(x^*)$. If $z^{(i)}$ and z^* have lower target densities than $x^{(i)}$ and x^* , respectively (a likely, but not required situation), then the acceptance probability in (3.13) reduces to $\min\{1, \pi(x^*)/\pi(x^{(i)})\}$, the acceptance probability of the Metropolis algorithm in (3.2). The proposed algorithm accepts the joint proposal (z^*, x^*) as $(z^{(i+1)}, x^{(i+1)})$ with the probability in (3.13) and sets $(z^{(i+1)}, x^{(i+1)})$ to $(z^{(i)}, x^{(i)})$ otherwise.

Altogether, each iteration of the RAM algorithm is composed of four steps as shown in Algorithm 1. The first three generate a proposal via three consecutive forced transitions; *Step 1* for the downward proposal x' given $x^{(i)}$, *Step 2* for the upward proposal x^* given x' , and *Step 3* for the downward proposal z^* given x^* . The last step determines whether the joint proposal, (z^*, x^*) , is accepted or not.

For computational efficiency, some density values in Algorithm 1 do not need to be calculated repeatedly. For example, $\pi(x')$ in *Step 2* is already evaluated during the forced

Algorithm 1. A repelling-attracting Metropolis algorithm.

Set initial values $z^{(0)}$ and $x^{(0)}$. For $i = 0, 1, \dots$

Step 1 : (\searrow) Resample $x' \sim q(x' | x^{(i)})$ and $u_1 \sim \text{Uniform}(0, 1)$ until $u_1 < \min\left\{1, \frac{\pi(x^{(i)})+\epsilon}{\pi(x')+\epsilon}\right\}$.

Step 2: (\nearrow) Resample $x^* \sim q(x^* | x')$ and $u_2 \sim \text{Uniform}(0, 1)$ until $u_2 < \min\left\{1, \frac{\pi(x^*)+\epsilon}{\pi(x')+\epsilon}\right\}$.

Step 3: (\searrow) Resample $z^* \sim q(z^* | x^*)$ and $u_3 \sim \text{Uniform}(0, 1)$ until $u_3 < \min\left\{1, \frac{\pi(x^*)+\epsilon}{\pi(z^*)+\epsilon}\right\}$.

Step 4: Set $(z^{(i+1)}, x^{(i+1)}) = (z^*, x^*)$ if $u_4 < \min\left\{1, \frac{\pi(x^*) \min\{1, (\pi(x^{(i)})+\epsilon)/(\pi(z^{(i)})+\epsilon)\}}{\pi(x^{(i)}) \min\{1, (\pi(x^*)+\epsilon)/(\pi(z^*)+\epsilon)\}}\right\}$,

where $u_4 \sim \text{Uniform}(0, 1)$, and set $(z^{(i+1)}, x^{(i+1)}) = (z^{(i)}, x^{(i)})$ otherwise.

downhill step in *Step 1*. Thus, if we save the value of $\pi(x')$ in *Step 1*, then we do not need to re-evaluate $\pi(x')$ in *Step 2*. Similarly, we need not re-evaluate $\pi(x^*)$ in *Step 3* if we save its value in *Step 2*. Also, $\pi(x^*)$ and $\pi(z^*)$ in *Step 4* are already evaluated in *Step 2* and *Step 3*, respectively, and thus they do not need to be re-calculated in *Step 4*. Lastly, since the density of the previous state $\pi(x^{(i)})$ is used both in *Step 1* and *Step 4*, it is better to evaluate and save this value before *Step 1* begins.

3.3 Numerical illustrations

We use a quad-core Intel Core i7 at 3.5 GHz and 16 GB of memory to run all the algorithms in the following examples. Both Metropolis and RAM algorithms are implemented under the same configuration, e.g., the same initial values and proposal scales, to show the benefit of simply replacing the Metropolis algorithm with the RAM algorithm.

3.3.1 Example 1: High-dimensional and multimodal distributions

The target distribution is a mixture of three d -dimensional Gaussian distributions;

$$\pi(x) = \frac{1}{4}N_d(x \mid -20 \times \mathbf{1}_d, I_d) + \frac{1}{2}N_d(x \mid 0 \times \mathbf{1}_d, I_d) + \frac{1}{4}N_d(x \mid 10 \times \mathbf{1}_d, I_d), \quad (3.14)$$

where $x = (x_1, x_2, \dots, x_d)^\top$, $\mathbf{1}_d$ is a vector of ones with length d , and I_d is a d -dimensional identity matrix. To increase difficulty, we assume that the leftmost mode at $-20 \times \mathbf{1}_d$ is unknown while the other two modes at $0 \times \mathbf{1}_d$ and $10 \times \mathbf{1}_d$ are known; the leftmost unknown mode is twice more distant from the central mode than the rightmost one. Here we investigate RAM's high dimensional behavior compared to that of a Metropolis algorithm in $d \in \{3, 10, 20\}$. For each dimension we implement RAM and Metropolis algorithms initialized at $z^{(0)} = 0 \times \mathbf{1}_d$ (only for RAM) and $x^{(0)} = 0 \times \mathbf{1}_d$ with $q(a \mid b) = N_d(a \mid b, \Sigma)$, where Σ is defined in (3.10).

In high dimension, an initial random-walk proposal with $S_0 = \sigma^2 I_d$ may result in a Markov chain that never jumps between modes during the burn-in period. In this case, using the

sample variance-covariance matrix S calculated from the burn-in samples being stuck at a local mode is unlikely to expedite jumping between modes after the burn-in period. To help a Markov chain visit at least the known modes during the burn-in period, we calculate S_0 as follows. We first run two short Markov chains using a random-walk Metropolis algorithm each of length 5,000 initialized at $0 \times \mathbf{1}_d$ and $10 \times \mathbf{1}_d$, respectively, whose scale matrix is $\sigma^2 I_d$, to make each chain be stuck at the known modes. The scale σ of this random-walk Metropolis algorithm is set to $2.38/\sqrt{d}$ for reasonable acceptance rate at around 0.3. Next, we calculate a sample variance-covariance matrix using the combined 10,000 samples and set it to the initial proposal scale matrix S_0 .

We run 20 chains each of length 100,000, discarding the first 50,000 samples as burn-ins. Because the RAM algorithm takes more CPU time than the Metropolis algorithm, we run longer chains for the Metropolis algorithm (still discarding the first 50,000 samples). For a fair comparison, we thin the longer chains to match the same sample size 50,000 each.

To evaluate each algorithm's jumping ability, we use three numerical measures. The first measure is the number of chains out of 20 that discover (visit) the unknown mode at $-20 \times \mathbf{1}_d$, denoted by N_{discover} . The second measure is the second largest eigenvalue of the transition matrix, denoted by λ_2 , based on the combined one million samples of the first coordinate, x_1 . The reason for considering only the first coordinate is that we can say a d -dimensional jump between modes occurs when a jump between modes occurs in the first coordinate, considering the target distribution in (3.14). To construct a transition matrix, we discretize the range of x_1 into three regions, $r_1 = \{x_1 : x_1 < -5\}$, $r_2 = \{x_1 : -5 \leq x_1 < 5\}$, and $r_3 = \{x_1 : 5 \leq x_1\}$. After constructing a 3×3 transition matrix based on r_1 , r_2 , and r_3 , we calculate the second largest eigenvalue of this matrix, λ_2 , which is related to the geometric convergence rate when the state-space is finite (Liu, 2008), i.e.,

$$\| P^n(\cdot | x^{(0)}) - \pi(\cdot) \|_{\text{TV}} \leq c\lambda_2^n,$$

where P^n is an n -step kernel distribution starting at $x^{(0)}$, $\| \cdot \|_{\text{TV}}$ is a total variation distance, c is a constant, and $|\lambda_2| < 1$. Thus, smaller λ_2 indicates the distribution of the samples

converges to the target distribution faster. Since λ_2 is related to a geometric convergence rate, even a small difference between two values of λ_2 can make a huge difference in terms of the convergence rate. The last measure is the number (proportion) of the jumps to the other modes among all the accepted proposals.

We summarize all the sampling results in Table 3.1, including the length of each chain before we discard the burn-in samples and thin each chain, mean CPU time averaged over 20 runs, acceptance rate, λ_2 , and N_{discover} ; the last two are calculated after we discard the burn-in samples and thin each chain for a fair comparison. The convergence rate of the RAM algorithm in terms of λ_2 is uniformly faster than that of the Metropolis algorithm, and the former discovers the unknown mode at $-20 \times \mathbf{1}_d$ better than the latter as dimensionality increases.

Using the combined one million samples (20 chains each of length 50,000) of the first coordinate x_1 obtained by each algorithm, we draw their histograms in Figure 3.2. The curves in the histograms represent the marginal target density of the first coordinate, i.e.,

$$\pi(x_1) = \frac{1}{4}\mathcal{N}_1(x_1 \mid -20, 1) + \frac{1}{2}\mathcal{N}_1(x_1 \mid 0, 1) + \frac{1}{4}\mathcal{N}_1(x_1 \mid 10, 1).$$

In dimension three, both algorithms recover the target distribution well. However, as dimensionality increases, i.e., $d \in \{10, 20\}$, the Metropolis algorithm deteriorates more quickly than

Table 3.1: Each chain’s sample size before we discard the burn-in samples and thin each chain, average CPU time in seconds, acceptance rate, the second largest eigenvalue of each transition matrix (λ_2), and the number of chains out of 20 that discover the unknown mode at $-20 \times \mathbf{1}_d$ (N_{discover}). Before we calculate λ_2 and N_{discover} , we discard the first 50,000 samples as burn-ins and thin each chain to match the same sample size 50,000 for a fair comparison.

| | $d = 3$ | | $d = 10$ | | $d = 20$ | |
|-----------------------|------------|---------|------------|---------|------------|---------|
| | Metropolis | RAM | Metropolis | RAM | Metropolis | RAM |
| Sample size | 411,419 | 100,000 | 571,635 | 100,000 | 712,366 | 100,000 |
| Average CPU time | 211 | 210 | 312 | 312 | 422 | 422 |
| Acceptance rate | 0.086 | 0.168 | 0.022 | 0.030 | 0.042 | 0.024 |
| λ_2 | 0.9392 | 0.9330 | 0.9999 | 0.9886 | 1.0000 | 0.9994 |
| N_{discover} | 20 | 20 | 6 | 20 | 2 | 8 |

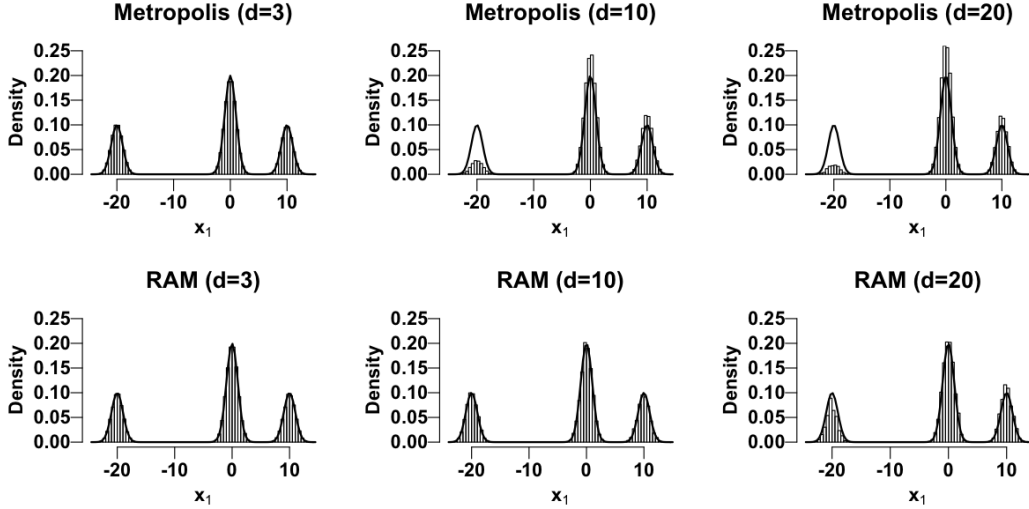


Figure 3.2: Histograms of the combined one million samples of the first coordinate (x_1) obtained by each algorithm in each dimension (d). The curve in each histogram represents the marginal target density function of x_1 .

the RAM algorithm in terms of restoring the target distribution because most Metropolis chains do not discover the unknown mode at $-20 \times \mathbf{1}_d$.

We also summarize the number of jumps to the other modes among all the accepted proposals in Table 3.2. The number of accepted proposals of the Metropolis algorithm in $d = 20$ is larger than that in $d = 10$ because more chains in $d = 20$ explore the known local modes without visiting the unknown mode. Clearly, RAM’s proportion of the accepted proposals that jump to the other modes is uniformly higher than that of the Metropolis algorithm in all dimensions.

Table 3.2: The number (#) and proportion of jumps to the other modes among the accepted proposals. We use the combined one million samples to calculate these.

| | $d = 3$ | | $d = 10$ | | $d = 20$ | |
|-------------------------|------------|---------|------------|--------|------------|--------|
| | Metropolis | RAM | Metropolis | RAM | Metropolis | RAM |
| # of accepted moves | 86,078 | 168,233 | 22,169 | 30,089 | 41,532 | 24,423 |
| # of jumps to the other | 23,352 | 75,321 | 2,236 | 13,530 | 346 | 1,647 |
| Proportion | 27.1% | 44.8% | 10.1% | 45.0% | 0.8% | 6.7% |

3.3.2 Example 2: A mixture of 20 bivariate Gaussian densities

For a comparison with other tempering methods, our second numerical illustration targets a mixture of 20 bivariate Gaussian distributions, given in Kou et al. (2006),

$$\pi(x) \propto \sum_{j=1}^{20} \frac{w_j}{\tau_j^2} \exp\left(-\frac{1}{2\tau_j^2}(x - \mu_j)^\top(x - \mu_j)\right),$$

where $x = (x_1, x_2)^\top$. The 20 mean vectors, $\{\mu_1, \dots, \mu_{20}\}$, are specified in Kou et al. (2006) and plotted in the first panel of Figure 3.3. Following Kou et al. (2006), we consider two cases; in case (a), the modes are equally weighted and have equal variances, $w_j = 1/20$ and $\tau_j^2 = 1/100$, and in case (b) weights and variances are unequal, $w_j = 1/\|\mu_j - (5, 5)^\top\|$ and $\tau_j^2 = \|\mu_j - (5, 5)^\top\|/20$. In case (b), modes near $(5, 5)$ are more weighted and have smaller variances. Contour plots of the target distributions in cases (a) and (b), respectively, appear in Figure 3.3. Contour lines correspond to 1%, 10%, 50%, and 95% probability.

Kou et al. (2006) used this target distribution to compare the equi-energy sampler and parallel tempering. We follow their simulation configurations by running the RAM algorithm for 100,000 iterations, discarding the first 50,000 for each of the two cases, i.e., the number of burn-ins, denoted by n_{burn} , is 50,000. The RAM algorithm is initialized at random values of $x^{(0)}$ and $z^{(0)}$ in the unit square, $[0, 1] \times [0, 1]$. We set $q(a | b) = N_2(a | b, \Sigma)$, where Σ is defined in (3.10); $S_0 = \sigma^2 I_2$ during the burn-in period and we calculate S , a sample variance-

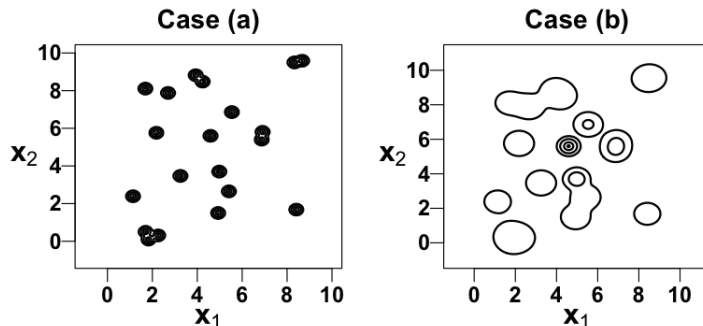


Figure 3.3: The first panel exhibits the contour plot of the target density in Example 3, case (a) and the second panel shows that of the target density in Example 3, case (b). The plotted contours outline regimes with probability 1%, 10%, 50%, and 95% under $\pi(x)$.

covariance matrix, using the first n_{burn} samples. We use an arbitrarily large proposal scale, e.g., $\sigma = 10$, pretending that the locations of the modes are unknown. The acceptance rates are 0.072 for case (a) and 0.293 for case (b).

Using the samples obtained by the RAM algorithm, we display the bivariate scatter plots for 50,000 samples, the bivariate trace plots for the last 2,000 iterations for case (a) and the last 1,000 iterations for case (b), and the autocorrelation functions for 50,000 samples of the first coordinate x_1 in Figure 3.4. The numbers of iterations used in the trace plots are the same as those in Kou et al. (2006). These plots can be compared to those for the equi-energy sampler and those for the parallel tempering provided in Kou et al. (2006).

To estimate moments, we again follow Kou et al. (2006) and run 20 independent chains using the RAM algorithm. Table 3.3 summarizes the moment estimates for each case, where results of the equi-energy sampler and parallel tempering are from Kou et al. (2006). The ratios of the mean squared error (MSE) of both the equi-energy sampler and parallel temper-

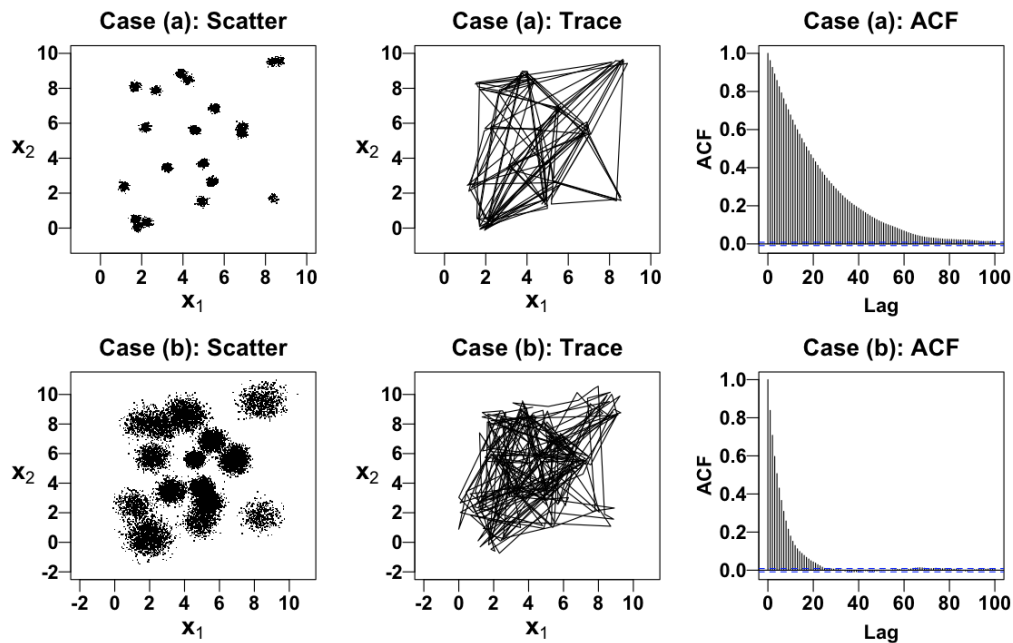


Figure 3.4: Results of the repelling-attracting Metropolis algorithm. The first column displays bivariate scatter plots for 50,000 samples, the middle column displays the bivariate trace plots for the last 2,000 samples for case (a) and the last 1,000 samples for case (b), and the last column displays the autocorrelation functions for 50,000 samples of x_1 .

Table 3.3: Moment estimates for cases (a) and (b) of Example 1 based on 20 independent chains, each of length 50,000, generated with the RAM algorithm, the equi-energy sampler (EE), and parallel tempering (PT). Results for the latter two samplers are reproduced from Kou et al. (2006). Estimates are the means over the 20 runs and standard deviations of the 20 runs are given in the parentheses next to estimates.

| Case (a) | Truth | RAM | EE | PT | MSE ratio (EE/RAM) | MSE ratio (PT/RAM) |
|------------|--------|-----------------|-----------------|-----------------|-----------------------|-----------------------|
| $E(x_1)$ | 4.478 | 4.5081 (0.081) | 4.5019 (0.107) | 4.4185 (0.170) | 1.61 | 4.34 |
| $E(x_2)$ | 4.905 | 4.8934 (0.097) | 4.9439 (0.139) | 4.8790 (0.283) | 2.18 | 8.46 |
| $E(x_1^2)$ | 25.605 | 25.9153 (0.843) | 25.9241 (1.098) | 24.9856 (1.713) | 1.62 | 4.11 |
| $E(x_2^2)$ | 33.920 | 33.8831 (0.952) | 34.4763 (1.373) | 33.5966 (2.867) | 2.42 | 9.17 |

| Case (b) | Truth | RAM | EE | PT | MSE ratio (EE/RAM) | MSE ratio (PT/RAM) |
|------------|--------|----------------|----------------|----------------|-----------------------|-----------------------|
| $E(x_1)$ | 4.688 | 4.693 (0.028) | 4.699 (0.072) | 4.709 (0.116) | 6.56 | 17.18 |
| $E(x_2)$ | 5.030 | 5.029 (0.031) | 5.037 (0.086) | 5.001 (0.134) | 7.74 | 19.54 |
| $E(x_1^2)$ | 25.558 | 25.742 (0.310) | 25.693 (0.739) | 25.813 (1.122) | 4.34 | 10.19 |
| $E(x_2^2)$ | 31.378 | 31.487 (0.347) | 31.433 (0.839) | 31.105 (1.186) | 5.34 | 11.20 |

ing to that of the RAM algorithm are greater than one, meaning that the RAM algorithm performs uniformly better than both in terms of MSE. The improvement is particularly striking for case (b) with unequal weights and variances.

However, we emphasize that this comparison does not take into account the CPU time, because the simulation configurations of Kou et al. (2006) does not account for different CPU time required by the equi-energy sampler or that required by the parallel tempering. The RAM algorithm, however, takes an average of 1,426 seconds in case (a) and 1,204 seconds in case (b), averaging over 20 independent runs.

3.3.3 Example 4: Time delay estimation problem

Our last numerical illustration targets a grossly multi-modal distribution whose modes are 600 standard deviations away from each other. This multi-modal distribution arises from an applied astrophysical project in Chapter 2 that originally motivated the RAM algorithm. Quasars are highly luminous astronomical sources in the distant Universe. If there is a massive galaxy between a quasar and the Earth, the gravitational field of the intervening

galaxy may act as a strong lens, bending the light rays emitted by the quasar. From our vantage points, two (or more) images of the quasar may appear in slightly different locations on the sky. This effect is known as strong gravitational lensing (Schneider et al., 2006). Because the light corresponding to the two images may take different routes to the Earth, their travel times may also differ. This difference is called a time delay. If we construct a time series of the brightness of each image, temporal features appear shifted in time between the two or more images because of the time delay. Accurate time delay estimation is important because it is, for example, used to calculate the current expansion rate of the Universe, i.e., the Hubble constant (Refsdal, 1964).

Figure 3.5 displays two irregularly-observed time series of the brightness of doubly-lensed quasar Q0957+561 (Hainline et al., 2012); the two time series are labeled A and B . Brightness is reported on a magnitude scale where smaller values correspond to brighter images. Let $x(t) \equiv \{x(t_1), \dots, x(t_n)\}$ and $y(t) \equiv \{y(t_1), \dots, y(t_n)\}$ denote the n observed magnitudes in time series A and B , respectively. Let $\delta(t) \equiv \{\delta(t_1), \dots, \delta(t_n)\}$ and $\eta(t) \equiv \{\eta(t_1), \dots, \eta(t_n)\}$ represent the n known standard deviations of the measurement errors for $x(t)$ and $y(t)$, respectively. There are 57 observations in the time series of Q0957+561, i.e., $n = 57$.

We assume that for each observed time series there is an unobserved underlying brightness curve. Let $X(t) \equiv \{X(t_1), \dots, X(t_n)\}$ denote the latent magnitudes for time series A and $Y(t) \equiv \{Y(t_1), \dots, Y(t_n)\}$ denote those for time series B . We assume that one of the latent brightness curves is a shifted version of the other, i.e.,

$$Y(t_j) = X(t_j - \Delta) + \beta_0, \tag{3.15}$$

where Δ is the unknown time delay and β_0 is an unknown magnitude offset. This is called a curve-shifted model.

Each observed magnitude is assumed to be independent Gaussian conditioning on its

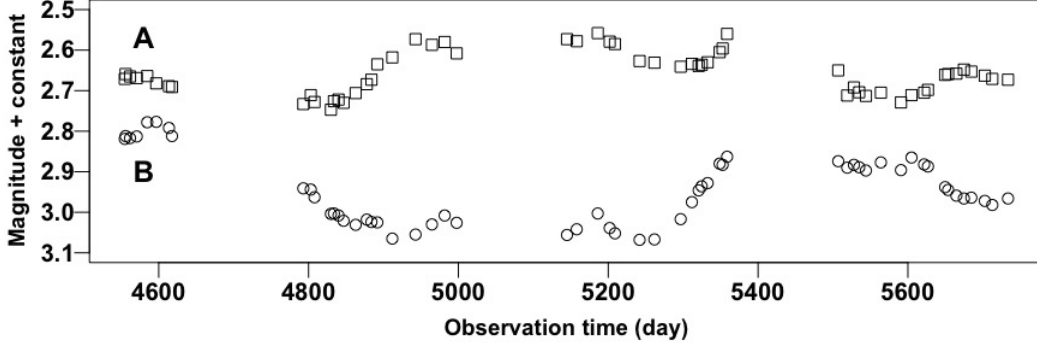


Figure 3.5: Two observed time series of doubly-lensed quasar Q0957+561 (Hainline et al., 2012). Time series A is denoted by squares and time series B is denoted by circles. Magnitude is an astronomical measure of brightness. Both time series are plotted with an offset (constant) in magnitude, but this does not affect the time delay estimation. Here we shifted time series B by 0.4 magnitude in the y -axis to display two time series in the same plot. The convention in astrophysics is to plot the magnitude inversely so that smaller magnitudes (brighter image) appear on the top and larger ones (fainter image) on the bottom.

latent magnitude,

$$\begin{aligned} x(t_j) | X(t_j) &\sim \text{Normal}(X(t_j), \delta^2(t_j)), \\ y(t_j) | Y(t_j) &\sim \text{Normal}(Y(t_j), \eta^2(t_j)). \end{aligned} \quad (3.16)$$

Using the model in (3.15), we can express (3.16) as

$$y(t_j) | X(t_j - \Delta), \Delta, \beta_0 \sim \text{Normal}(X(t_j - \Delta) + \beta_0, \eta^2(t_j)).$$

We assume that the latent magnitudes follow an Ornstein-Uhlenbeck process (Kelly et al., 2009). The solution of a stochastic differential equation of the Ornstein-Uhlenbeck process yields the sampling distribution of the time-sorted latent magnitudes $X(t^\Delta)$, where $t^\Delta \equiv (t_1^\Delta, \dots, t_{2n}^\Delta)^\top$ is the sorted $2n$ times among the n observation times, t , and the n time-delay-shifted observation times, $t - \Delta$. Specifically,

$$\begin{aligned} X(t_1^\Delta) | \Delta, \theta &\sim \text{Normal}\left(\mu, \frac{\tau\phi^2}{2}\right), \quad \text{and for } j = 2, 3, \dots, 2n, \\ X(t_j^\Delta) | X(t_{j-1}^\Delta), \Delta, \theta &\sim \text{Normal}\left(\mu + a_j(X(t_{j-1}^\Delta) - \mu), \frac{\tau\phi^2}{2}(1 - a_j^2)\right), \end{aligned}$$

where $\theta \equiv (\mu, \phi^2, \tau)^\top$ and $a_j = \exp(-(t_j^\Delta - t_{j-1}^\Delta)/\tau)$.

A Bayesian analysis requires prior distributions on the several parameters; Δ follows a Uniform $[t_1 - t_n, t_n - t_1] = [-1178.939, 1178.939]$, β_0 follows a Uniform $[-60, 60]$, μ follows a Uniform $[-30, 30]$, ϕ^2 follows an inverse-Gamma $(1, 2/10^7)$, and τ follows an inverse-Gamma $(1, 1)$, where a density function of $v \sim \text{inverse-Gamma}(a, b)$ is proportional to $v^{-a-1} \exp(-b/v)$. Further details and motivation for this model, including the choice of prior distributions, are given in Chapter 2.

To sample from the joint posterior density function, $\pi(X(t^\Delta), \Delta, \beta_0, \theta \mid x(t), y(t))$, we adopt a Metropolis-Hastings within Gibbs sampler (MHwG, Tierney, 1994b) composed of three steps as shown in Algorithm 2 below. We suppress conditioning on $x(t)$ and $y(t)$ in all three steps here and elsewhere. The factorization in *Step 1* means that we first sample Δ given β_0 and θ , and then sample $X(t^\Delta)$ given Δ , β_0 , and θ . See Appendix B for details of the necessary complete conditional distributions.

Because the marginal posterior distribution of the time delay is often multimodal, we compare a Metropolis algorithm and tempered transitions (Neal, 1996) with a RAM algorithm to sample Δ from $\pi(\Delta \mid \beta_0, \theta)$ in *Step 1* of Algorithm 2.

At each iteration, the tempered transitions ascend (heating) the temperature ladder to explore a flatter surface where the modes are melted down, and then descend (cooling) the ladder, accepting the last candidate with a modified acceptance probability to maintain the stationary distribution. Specifically, suppose $\pi_j(\Delta) \propto \{\pi(\Delta \mid \beta_0^{(i-1)}, \theta^{(i-1)})\}^{1/T_j}$, where T_j is the temperature at rung j of the temperature ladder, for $j = 1, \dots, J$. The target density is $\pi_0(\Delta)$ with $T_0 = 1$ and the ladder has J rungs with $T_0 = 1 < T_1 < \dots < T_J$. At

Algorithm 2. A Metropolis-Hastings within Gibbs sampler for the time delay model.

Set initial values $\Delta^{(0)}$, $X^{(0)}(t^{\Delta^{(0)}})$, $\beta_0^{(0)}$, and $\theta^{(0)}$. For $i = 1, 2, \dots$

$$\begin{aligned} \text{Step 1: Draw } \left(X^{(i)}(t^{\Delta^{(i)}}), \Delta^{(i)} \right) &\sim \pi \left(X(t^\Delta), \Delta \mid \beta_0^{(i-1)}, \theta^{(i-1)} \right) \\ &= \pi \left(X(t^\Delta) \mid \Delta, \beta_0^{(i-1)}, \theta^{(i-1)} \right) \pi \left(\Delta \mid \beta_0^{(i-1)}, \theta^{(i-1)} \right). \end{aligned}$$

$$\text{Step 2: Draw } \beta_0^{(i)} \sim \pi \left(\beta_0 \mid \theta^{(i-1)}, X^{(i)}(t^{\Delta^{(i)}}), \Delta^{(i)} \right).$$

$$\text{Step 3: Draw } \theta^{(i)} \sim \pi \left(\theta \mid X^{(i)}(t^{\Delta^{(i)}}), \Delta^{(i)}, \beta_0^{(i)} \right).$$

the beginning of iteration i , we generate $\hat{\Delta}_1$ from $\text{Normal}(\Delta^{(i-1)}, \sigma_1^2)$, and accept it with probability $\min(1, \pi_1(\hat{\Delta}_1)/\pi_1(\Delta^{(i-1)}))$ and set $\hat{\Delta}_1 = \Delta^{(i-1)}$ otherwise. Next, we generate $\hat{\Delta}_2$ from $\text{Normal}(\hat{\Delta}_1, \sigma_2^2)$, and accept it with probability $\min(1, \pi_2(\hat{\Delta}_2)/\pi_2(\hat{\Delta}_1))$ and set $\hat{\Delta}_2 = \hat{\Delta}_1$ otherwise. We repeat this process until we reach the top of the temperature ladder, collecting $\hat{\Delta}_1, \dots, \hat{\Delta}_J$. At the top, we generate $\check{\Delta}_{J-1}$ from $\text{Normal}(\hat{\Delta}_J, \sigma_J^2)$, and accept it with probability $\min(1, \pi_{J-1}(\check{\Delta}_{J-1})/\pi_{J-1}(\hat{\Delta}_J))$ and set $\check{\Delta}_{J-1} = \hat{\Delta}_J$ otherwise. We repeat this process until we reach the bottom of the temperature ladder, collecting $\check{\Delta}_{J-1}, \dots, \check{\Delta}_0$. We set $\Delta^{(i)} = \check{\Delta}_0$ with probability

$$\min \left\{ 1, \frac{\pi_1(\Delta^{(i-1)})}{\pi_0(\Delta^{(i-1)})} \times \dots \times \frac{\pi_J(\hat{\Delta}_{J-1})}{\pi_{J-1}(\hat{\Delta}_{J-1})} \frac{\pi_{J-1}(\check{\Delta}_{J-1})}{\pi_J(\check{\Delta}_{J-1})} \times \dots \times \frac{\pi_0(\check{\Delta}_0)}{\pi_1(\check{\Delta}_0)} \right\}$$

and set $\Delta^{(i)} = \Delta^{(i-1)}$ otherwise.

To sample Δ from $\pi(\Delta | \beta_0, \theta)$ via the RAM algorithm, we additionally keep track of the auxiliary variable during the run, i.e., $\{z^{(i)}, i = 0, 1, 2, \dots\}$. At iteration i , we sequentially draw $\Delta' \sim q^D(\Delta' | \Delta^{(i-1)})$, $\Delta^* \sim q^U(\Delta^* | \Delta')$, and $z^* \sim q^D(z^* | \Delta^*)$. We set $(z^{(i)}, \Delta^{(i)})$ to (z^*, Δ^*) with probability $\alpha^J(z^*, \Delta^* | z^{(i-1)}, \Delta^{(i-1)})$ given in (3.13), and set $(z^{(i)}, \Delta^{(i)})$ to $(z^{(i-1)}, \Delta^{(i-1)})$ otherwise. Because $\{z^{(i)}, i = 0, 1, 2, \dots\}$ are introduced solely to enable sampling Δ from $\pi(\Delta | \beta_0, \theta)$, only $\Delta^{(i)}$ is used to sample $X(t^\Delta)$, β_0 , and θ for the following steps in Algorithm 2, and $z^{(i)}$ is used to draw $\Delta^{(i+1)}$ at the next iteration.

We fit the time delay model using the MHwG sampler equipped with a Metropolis algorithm, tempered transitions, or a RAM algorithm. In each case, we independently run five chains each of length 150,000, discarding the first 50,000. All algorithms are initialized at the same point; $z^{(0)} = 0$ (only for RAM), $\beta_0^{(0)} = \sum_{j=1}^n \{y(t_j) - x(t_j)\}/n = -0.113$, $\mu^{(0)} = \sum_{j=1}^n x(t_j)/n = 2.658$, $\phi^{(0)} = 0.01$, $\tau^{(0)} = 200$, and $X^{(0)}(t^{\Delta^{(0)}})$ is a vector of $x(t)$ and $y(t - \Delta^{(0)}) - \beta_0^{(0)}$ that are sorted in time. When it comes to the initial value of the time delay, $\Delta^{(0)}$, we spread five initial values, $\{-1000, -500, 0, 500, 1000\}$, across the five chains as is commonly done to check the multimodal behavior of Δ .

For Metropolis and RAM algorithms, we set $q(a | b) = N_1(a | b, \Sigma)$, where Σ is defined in (3.10). Because we do not know the information about the locations of the modes, we

set an arbitrarily large initial proposal scale $\sigma = 500$ ($S_0 = \sigma^2$) during the burn-in period, about a quarter of the length of the entire range of Δ . After the burn-in period, we calculate a sample standard deviation from all the posterior samples of Δ drawn during the burn-in period and set it to σ ($S = \sigma^2$).

Tempered transitions require several tuning parameters, i.e., the number of rungs of the temperature ladder, the temperature of each rung, and the proposal scales. Setting these parameters can be challenging in practice (Behrens et al., 2012). To fit the Q0957+561 data, we set five rungs ($J = 5$) with corresponding temperature $T_j = 4^j$ and proposal scale $\sigma_j = \sigma \times 1.2^{j-1}$ for $j = 1, \dots, 5$. The common proposal scale on each rung σ plays the same role as that in the Metropolis and RAM algorithms, and thus we set $\sigma = 500$ as an arbitrarily large initial proposal scale. After the burn-in period, we calculate the sample standard deviation of the posterior samples of Δ and set it to σ .

Considering the different CPU time taken for each algorithm, we run longer chains of the MHwG equipped with the Metropolis and RAM algorithms and thin these chains to match the same sample size 100,000 for each chain. Table 3.4 summarizes all the sampling results including the sample size of each chain before we discard the burn-in samples and thin each chain, average CPU time over five runs, acceptance rate for Δ , and the number of jumps between two distant modes near 400 days and 1,100 days, respectively (N_{jumps}).

For all algorithms, two chains out of five have discovered the mode near 400 days and

Table 3.4: Each chain’s sample size before we discard the 50,000 burn-in samples and thin each chain, average CPU time over five runs in seconds, acceptance rate for Δ , and the number of jumps between the grossly distant modes near 400 days and 1,100 days, respectively (N_{jumps}). Before calculating N_{jumps} , we discard the first 50,000 samples as burn-ins and thin each chain to match the same sample size 100,000 for a fair comparison.

| | Sample size | Average CPU time | Acceptance rate | N_{jumps} |
|----------------------|-------------|------------------|-----------------|--------------------|
| Tempered transitions | 150,000 | 3,058 | 0.203 | 17 |
| Metropolis | 1,776,813 | 3,060 | 0.228 | 26 |
| RAM | 492,216 | 3,057 | 0.292 | 72 |

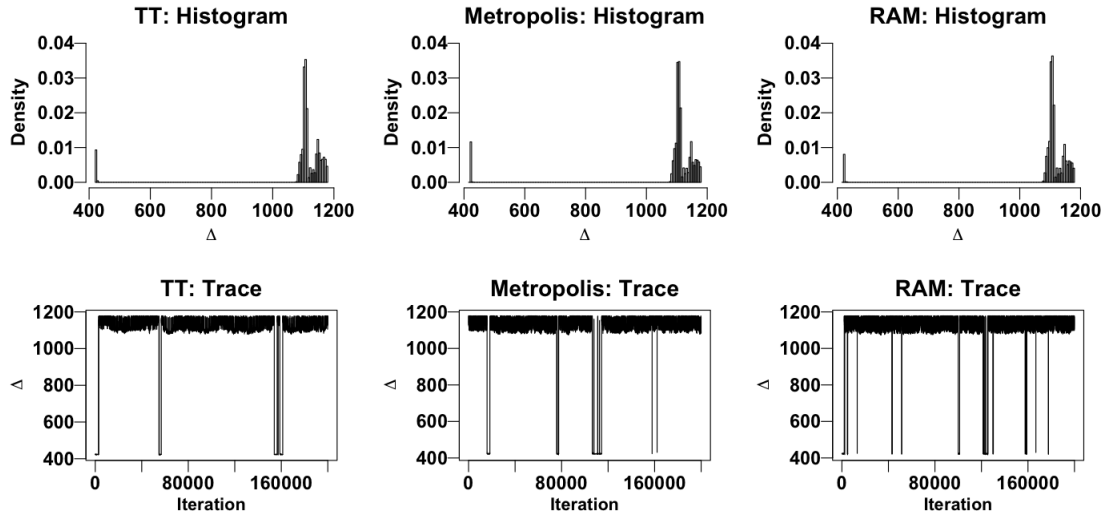


Figure 3.6: The histograms and trace plots of 200,000 samples of Δ , based on the two chains that have discovered the mode near 400 days, obtained by tempered transitions (TT) appear in the first column, those obtained by the Metropolis algorithm appear in the second column, and those obtained by the RAM algorithm in the third column.

the other three chains have been stuck at the modes near 1,100 days. In these two chains, the numbers of jumps between the modes for tempered transitions are $\{11, 6\}$, summing to 17. Similarly, the numbers of jumps for the Metropolis algorithm are $\{10, 16\}$, summing to 26, and those for the RAM algorithm are $\{23, 49\}$, summing to 72. This means that the total number of jumps per unit CPU time for the RAM algorithm is 2.77 times larger than that for the Metropolis algorithm and 4.24 times larger than that for tempered transitions.

3.4 Conclusion

A Metropolis algorithm is widely used due to its simplicity, though it is known to be inappropriate for exploring a multimodal distribution. To improve its ability to explore a multimodal distribution, we propose a repelling-attracting Metropolis (RAM) algorithm that can be implemented with a single tuning parameter like a Metropolis algorithm. Thus, the RAM algorithm can be an immediate alternative when users realize their Metropolis algorithm does not explore a multimodal distribution well. Its simple implementation can be appeal-

ing to both statisticians and practitioners because most temperature-based methods may require significant human time for tuning, especially for non-experts.

We do not believe, however, that the RAM algorithm will always perform more favorably than the tempering-based methods, and more work needs to be done to extend its applicability. In particular, we need to compare the theoretical convergence rate of our algorithm to others, though the intractable down-up proposal density can be a challenge for this purpose. Also, different ways to encourage a down-up movement in density may exist, e.g., mixing anti-Langevin and Langevin algorithms as suggested by Christian P. Robert or tempering with negative and positive temperature levels as suggested by Art B. Owen. Another avenue for further improvement would be allowing an asymmetric density function q so that a downhill move reaches out further than an uphill move does. Furthermore, it may be possible to generalize our method to handle a case where π itself is intractable. Applying this down-up idea to finding a global optimum of a multimodal density function is another possible extension as the tempering idea is used for a statistical annealing. We leave these for future research.

Appendix A

Proofs of Theorem, Lemma, and Corollary in Chapter 1

A.1 Proof of Lemma 1.3.1

If group j is interior ($1 \leq y_j \leq n_j - 1$, $n_j \geq 2$), we can derive an upper bound for the Beta-Binomial probability mass function of interior group j with respect to r and $\boldsymbol{\beta}$ as follows. All bounds in this proof are up to a constant multiple. With notation $q_j^E = 1 - p_j^E$,

$$\pi_{\text{obs}}(y_j \mid r, \boldsymbol{\beta}) \propto \frac{B(y_j + rp_j^E, n_j - y_j + rq_j^E)}{B(rp_j^E, rq_j^E)} \quad (\text{A.1})$$

$$= \frac{B(1 + rp_j^E, 1 + rq_j^E)}{B(rp_j^E, rq_j^E)} \frac{B(y_j + rp_j^E, n_j - y_j + rq_j^E)}{B(1 + rp_j^E, 1 + rq_j^E)} \quad (\text{A.2})$$

$$= \frac{rp_j^E q_j^E}{1 + r} \frac{B(y_j + rp_j^E, n_j - y_j + rq_j^E)}{B(1 + rp_j^E, 1 + rq_j^E)} \quad (\text{A.3})$$

$$= \frac{rp_j^E q_j^E}{1 + r} \frac{\int_0^1 v^{y_j - 1 + rp_j^E} (1 - v)^{n_j - y_j - 1 + rq_j^E} dv}{\int_0^1 v^{rp_j^E} (1 - v)^{rq_j^E} dv} \leq \frac{rp_j^E q_j^E}{1 + r}. \quad (\text{A.4})$$

The ratio of the two beta functions in (A.4) is less than or equal to one because the integrand of the beta function in the numerator is less than or equal to the integrand of the beta function in the denominator, considering that $0 \leq y_j - 1 \leq n_j - 2$ and $0 \leq n_j - y_j - 1 \leq n_j - 2$.

A lower bound for the ratio of the two beta functions in (A.1) is

$$\frac{B(y_j + rp_j^E, n_j - y_j + r(1 - p_j^E))}{B(rp_j^E, r(1 - p_j^E))} \quad (\text{A.5})$$

$$= \frac{(y_j - 1 + rp_j^E) \cdots (1 + rp_j^E) rp_j^E (n_j - y_j - 1 + rq_j^E) \cdots (1 + rq_j^E) rq_j^E}{(n_j - 1 + r)(n_j - 2 + r) \cdots (1 + r)r} \quad (\text{A.6})$$

$$\geq \frac{r^2 p_j^E q_j^E}{(n_j - 1 + r)(n_j - 2 + r) \cdots (1 + r)r} \geq \frac{rp_j^E q_j^E}{(n_{\max} + r)^{n_j - 1}} \geq \frac{rp_j^E q_j^E}{(1 + r)^{n_j - 1}} \quad (\text{A.7})$$

where $n_{\max} \equiv \max\{n_1, \dots, n_k\}$. The first inequality in (A.7) holds because each factor (except rp_j^E and rq_j^E) in the numerator of (A.6) is greater than or equal to one. The third inequality holds up to a constant multiple, $1/n_j^{n_j - 1}$, because $(n_{\max} + r)/(1 + r) \leq n_j$.

If group j is extreme with all successes ($y_j = n_j \geq 1$), the upper bound for the Beta-Binomial probability mass function of group j with respect to r and $\boldsymbol{\beta}$ is

$$\pi_{\text{obs}}(y_j = n_j \mid r, \boldsymbol{\beta}) \propto \frac{B(n_j + rp_j^E, rq_j^E)}{B(rp_j^E, rq_j^E)} \leq \frac{B(1 + rp_j^E, rq_j^E)}{B(rp_j^E, rq_j^E)} = p_j^E. \quad (\text{A.8})$$

The inequality holds because the integrand of the beta function in the numerator becomes the largest when $n_j = 1$. The lower bound for the Beta-Binomial probability mass function of this extreme group with respect to r and $\boldsymbol{\beta}$ is

$$\frac{B(n_j + rp_j^E, rq_j^E)}{B(rp_j^E, rq_j^E)} = \frac{(n_j - 1 + rp_j^E)(n_j - 2 + rp_j^E) \cdots (1 + rp_j^E) p_j^E}{(n_j - 1 + r)(n_j - 2 + r) \cdots (1 + r)} \geq (p_j^E)^{n_j}. \quad (\text{A.9})$$

The inequality holds because the ratio of the two beta functions in (A.9) is a decreasing function of r , and thus the lower bound is achieved as r goes to infinity.

Similarly, when group j is extreme with all failures ($y_j = 0, n_j \geq 1$), we can bound the ratio of the two beta functions of this extreme group by

$$(q_j^E)^{n_j} < \frac{B(rp_j^E, n_j + rq_j^E)}{B(rp_j^E, rq_j^E)} < q_j^E. \quad (\text{A.10})$$

A.2 Proof of Theorem 1.3.1

Because the r part of the upper bound for $L(r, \boldsymbol{\beta})$ in Lemma 1.3.2, i.e., $r^k/(1 + r)^k$, is always less than one, an upper bound for $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$, up to a normalizing constant, factors

into a function of r and a function of $\boldsymbol{\beta}$ as follows:

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y}) \propto f(r)g(\boldsymbol{\beta})L(r, \boldsymbol{\beta}) < f(r) \times \prod_{j=1}^k p_j^E q_j^E. \quad (\text{A.11})$$

The integration of $f(r)$ with respect to r is finite because it is a proper hyper-prior PDF. The integration of $\prod_{j=1}^k p_j^E q_j^E$ with respect to $\boldsymbol{\beta}$ is finite if and only if the covariate matrix of all groups, X , is of full rank m . To show the sufficient condition, let us choose m sub-groups, whose index set is denoted by W_{sub} , such that the $m \times m$ covariate matrix of the sub-groups is still of full rank m . Then,

$$\prod_{j=1}^k p_j^E q_j^E < \prod_{j \in W_{\text{sub}}} p_j^E q_j^E = \prod_{j \in W_{\text{sub}}} \frac{\exp(\mathbf{x}_j^\top \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})]^2}. \quad (\text{A.12})$$

The integration of this upper bound in (A.12) with respect to $\boldsymbol{\beta}$ factors into m separate integrations after linear transformations, $h_j = \mathbf{x}_j^\top \boldsymbol{\beta}$ for all $j \in W_{\text{sub}}$, whose Jacobian is a constant:

$$\int_{R^m} \prod_{j \in W_{\text{sub}}} \frac{\exp(\mathbf{x}_j^\top \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})]^2} d\boldsymbol{\beta} \propto \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j \in W_{\text{sub}}} \frac{\exp(h_j)}{[1 + \exp(h_j)]^2} dh_j = 1. \quad (\text{A.13})$$

Each integration on the right hand side leads to one because each integrand is a proper density function of the standard logistic distribution with respect to h_j .

Next, we show that if the rank of X is not of full rank m , then the integration of the $\boldsymbol{\beta}$ part of the lower bound for $L(r, \boldsymbol{\beta})$ in Lemma 1.3.2, i.e., $\prod_{j=1}^k p_j^E q_j^E$, cannot be finite. Without loss of generality, let us assume that the rank of X is $m - 1$ and that the last column of X can be expressed as a linear function of the first $m - 1$ columns. Due to the singularity of X , we can always find $m - 1$ linear functions, $t_i(\beta_i, \beta_m)$, $i = 1, 2, \dots, m - 1$, such that $\mathbf{x}_j^\top \boldsymbol{\beta} = x_{j1}t_1(\beta_1, \beta_m) + x_{j2}t_2(\beta_2, \beta_m) + \cdots + x_{j,m-1}t_{m-1}(\beta_{m-1}, \beta_m)$. As a result, the integration of $\prod_{j=1}^k p_j^E q_j^E$ with respect to $\boldsymbol{\beta}$ is infinity after a linear transformation from $\boldsymbol{\beta}$ to $(\beta_1^* = t_1(\beta_1, \beta_m), \beta_2^* = t_2(\beta_2, \beta_m), \dots, \beta_{m-1}^* = t_{m-1}(\beta_{m-1}, \beta_m), \beta_m)^\top$, whose Jacobian is one. For notational simplicity, we use two $(m - 1) \times 1$ vectors, $\mathbf{x}_j^* \equiv (x_{j1}, x_{j2}, \dots, x_{j,m-1})^\top$ and $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \dots, \beta_{m-1}^*)^\top$:

$$\int_{R^m} \prod_{j=1}^k \frac{\exp(\mathbf{x}_j^\top \boldsymbol{\beta})}{[1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})]^2} d\boldsymbol{\beta} = \int_{R^{m-1}} \prod_{j=1}^k \frac{\exp(\mathbf{x}_j^{*T} \boldsymbol{\beta}^*)}{[1 + \exp(\mathbf{x}_j^{*T} \boldsymbol{\beta}^*)]^2} d\boldsymbol{\beta}^* \times \int_R d\beta_m, \quad (\text{A.14})$$

where $\int_R d\beta_m = \infty$.

A.3 Proof of Corollary 1.3.1

Regarding the sufficient conditions for posterior propriety, an upper bound for $L(r, \boldsymbol{\beta})$ up to a constant multiplication is

$$L(r, \boldsymbol{\beta}) \propto \prod_{j=1}^k \frac{B(y_j + rp_j^E, n_j - y_j + rq_j^E)}{B(rp_j^E, rq_j^E)} < \prod_{j \in W_y} \frac{B(y_j + rp_j^E, n_j - y_j + rq_j^E)}{B(rp_j^E, rq_j^E)} \quad (\text{A.15})$$

$$= \prod_{j \in W_y} \frac{rp_j^E q_j^E}{1+r} \frac{B(y_j + rp_j^E, n_j - y_j + rq_j^E)}{B(1 + rp_j^E, 1 + rq_j^E)} \leq \frac{r^{k_y} \prod_{j \in W_y} p_j^E q_j^E}{(1+r)^{k_y}}. \quad (\text{A.16})$$

The inequality in (A.15) holds because the upper bound for the ratio of two beta functions for extreme group j is either $p_j^E (< 1)$ in (A.8) or $q_j^E (< 1)$ in (A.10). The inequality in (A.16) holds because the integrand of the beta function in the numerator is less than or equal to the integrand of the beta function in the denominator.

The upper bound for $L(r, \boldsymbol{\beta})$ in (A.16) would be the same as the upper bound for $L(r, \boldsymbol{\beta})$ in Lemma 1.3.2 if we removed all extreme groups from the data and treated the interior groups as a new data set ($k_y = k$). Thus, if the joint posterior density function $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is proper with the new data set of k_y interior groups based on Theorem 1.3.1, 1.3.2, or 1.3.3, then posterior propriety with the original data with all interior and all extreme groups combined ($1 \leq k_y \leq k - 1$) also holds. In other words, the extreme groups do not affect the sufficient condition for posterior propriety no matter how many of them are in the data as long as there exists at least one interior group in the data.

For the necessary conditions for posterior propriety, we will show that if a new data set with all the extreme groups removed does not meet the conditions for posterior propriety based on Theorem 1.3.1, 1.3.2, or 1.3.3, then $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is still improper even after we add extreme groups into the new data.

Because a lower bound for the Beta-Binomial probability mass function for extreme group j is either $(p_j^E)^{n_j}$ in (A.9) or $(q_j^E)^{n_j}$ in (A.10), the extra product term for extreme

groups to the lower bound for the likelihood function based only on interior groups is $\prod_{i \in W_y^c} (p_i^E)^{n_i I_{\{y_i=n_i\}}} (q_i^E)^{n_i I_{\{y_i=0\}}}$.

Specifically, let us consider a proper hyper-prior PDF for r , $f(r)$, and an improper flat hyper-prior PDF for $\boldsymbol{\beta}$, $g(\boldsymbol{\beta}) \propto d\boldsymbol{\beta}$ as in Theorem 1.3.1. Suppose we removed all the extreme groups in the data. If the rank of X_y is not of full rank, e.g., $\text{rank}(X_y) = m - 1$, then we see the term $\int_R d\beta_m$ in (A.14). This term does not disappear even after we add all the extreme groups to the data because multiplying $\prod_{i \in W_y^c} (p_i^E)^{n_i I_{\{y_i=n_i\}}} (q_i^E)^{n_i I_{\{y_i=0\}}}$ by the first integrand in (A.14) cannot make the term, $\int_R d\beta_m$, disappear. It means that $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is still improper.

Next, we consider $f(r) \propto dr/r^{u+1}$ for positive u and a proper hyper-prior PDF on $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, as in Theorem 1.3.2. Because contribution of extreme groups to the lower bound for the likelihood function, i.e., $\prod_{i \in W_y^c} (p_i^E)^{n_i I_{\{y_i=n_i\}}} (q_i^E)^{n_i I_{\{y_i=0\}}}$, is free of r , if k_y is smaller than $u + 1$, then $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is still improper even after we add all the extreme groups into the data.

If the data of interior groups do not meet the condition for posterior propriety specified in Theorem 1.3.3, then adding the extreme groups cannot change the result of posterior propriety. This is because Theorem 1.3.3 is an improper mixture of Theorem 1.3.1 and 1.3.2 and we already showed that extreme groups can be ignored in determining posterior propriety in Theorem 1.3.1 and 1.3.2.

A.4 Proof of Theorem 1.3.4

Considering the upper bound of the likelihood function in (1.16) when all groups are extreme ($k_y = 0$), the upper bound of $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ up to a constant multiple is

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y}) \propto f(r)g(\boldsymbol{\beta})L(r, \boldsymbol{\beta}) \leq f(r) \prod_{j=1}^k (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}}. \quad (\text{A.17})$$

The integration of $f(r)$ with respect to r is finite because $f(r)$ is proper. The integration of the $\boldsymbol{\beta}$ part in (A.17), i.e.,

$$\prod_{j=1}^k (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}}, \quad (\text{A.18})$$

with respect to $\boldsymbol{\beta}$ is finite if there exists a finite value of $\boldsymbol{\beta}$ that maximizes (A.18). This is because (A.18) is essentially a likelihood function of a logistic regression in (1.8) in that the powers in (A.18) are either one or zero with $I_{\{y_j=0\}} = 1 - I_{\{y_j=n_j\}}$. Thus, we can use the fact that the posterior distribution of $\boldsymbol{\beta}$ with its constant prior (Lebesgue measure) in a logistic regression is proper if there exists a finite MLE of $\boldsymbol{\beta}$ (Albert and Anderson, 1984; Speckman et al., 2009). (Jacobsen (1989) shows that the MLE of a logistic regression is unique if it exists.) Consequently, the integration of (A.18) with respect to $\boldsymbol{\beta}$ is finite if there exists a finite value of $\boldsymbol{\beta}$ that maximizes (A.18).

The lower bound of $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ up to a constant multiple can be derived from the lower bound of the likelihood function in (1.16), i.e.,

$$\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y}) \propto f(r)g(\boldsymbol{\beta})L(r, \boldsymbol{\beta}) \geq f(r) \prod_{j=1}^k \left[(p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \right]^{n_j}. \quad (\text{A.19})$$

The integration of the $\boldsymbol{\beta}$ part in (A.19) with respect to $\boldsymbol{\beta}$ can be bounded from below by

$$\begin{aligned} \int_{\mathbf{R}^m} \prod_{j=1}^k \left[(p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \right]^{n_j} d\boldsymbol{\beta} &\geq \int_{\mathbf{R}^m} \left[\prod_{j=1}^k (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \right]^{n_{\max}} d\boldsymbol{\beta} \\ &\geq \left[\int_{\mathbf{R}^m} \prod_{j=1}^k (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} d\boldsymbol{\beta} \right]^{n_{\max}}, \end{aligned} \quad (\text{A.20})$$

where the first inequality holds because of the largest power $n_{\max} \equiv \max(n_1, n_2, \dots, n_k)$ and the second inequality holds via Jensen's inequality because the power function is convex. The integrand in (A.20) is the same as (A.18). This indicates that the integration in (A.20) is not finite (and thus $\pi_{\text{hyp.post}}(r, \boldsymbol{\beta} \mid \mathbf{y})$ is improper) if a finite value of $\boldsymbol{\beta}$ that maximizes (A.18) does not exist (Albert and Anderson, 1984; Speckman et al., 2009).

A.5 Proof of Theorem 1.3.5

First, we show that the integration of (A.18) with respect to $\boldsymbol{\beta}$ is finite if (i) there are at least m clusters of groups whose covariate values are the same within each cluster and different between clusters, and (ii) in each cluster there are at least one group of all successes and at least one group of all failures. We define c_i as the index set of cluster i , e.g., $c_i = \{2, 5\}$ means that groups 2 and 5 are in cluster i . Then we can bound (A.18) with groups only in the m clusters as follows.

$$\prod_{j=1}^k (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \leq \prod_{j \in \{c_i, i=1,2,\dots,m\}} (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}} \leq \prod_{i=1}^m p_{c_i}^E q_{c_i}^E, \quad (\text{A.21})$$

where $p_{c_i}^E = 1 - q_{c_i}^E = \exp(\mathbf{x}_{c_i}^\top \boldsymbol{\beta}) / \{1 + \exp(\mathbf{x}_{c_i}^\top \boldsymbol{\beta})\}$ is the same expected random effect for all groups in cluster i and \mathbf{x}_{c_i} is the same covariate vector for all groups in cluster i . The first equality holds because some groups may not be included in one of m clusters. The second inequality holds for two reasons. First, groups in the same cluster share the same covariate values, meaning that every group in cluster i has the same expected random effect, $p_{c_i}^E = 1 - q_{c_i}^E$. Second, in each cluster there are at least one group with all successes and at least one group with all failures, indicating that in cluster i , $p_{c_i}^E q_{c_i}^E$ is the largest value of $\prod_{j \in c_i} (p_j^E)^{I_{\{y_j=n_j\}}} (q_j^E)^{I_{\{y_j=0\}}}$. The integration of the upper bound in (A.21) is finite with a linear transformation, $h_i = \mathbf{x}_{c_i}^\top \boldsymbol{\beta}$, as follows:

$$\int_{R^m} \prod_{i=1}^m p_{c_i}^E q_{c_i}^E d\boldsymbol{\beta} \propto \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \prod_{i=1}^m \frac{\exp(h_i)}{[1 + \exp(h_i)]^2} dh_i = 1. \quad (\text{A.22})$$

The last equality holds because $\exp(h_i)/[1 + \exp(h_i)]^2$ is a PDF of a standard Logistic distribution with respect to h_i .

These conditions also become necessary conditions when $\mathbf{x}_j^\top \boldsymbol{\beta} = \beta_1$ for all j . In this case, the conditions simply reduce to having at least one group with all successes and at least one group with all failures. Let us use notation $p_j^E = p^E = 1 - q^E = \exp(\beta_1)/(1 + \exp(\beta_1))$. If all the extreme groups have only successes ($y_j = n_j$ for all j), then we can bound $\pi_{\text{hyp.post}}(r, \beta_1 |$

\mathbf{y}) from below using the lower bound in (1.16) up to a normalizing constant as follows:

$$\pi_{\text{hyp.post}}(r, \beta_1 \mid \mathbf{y}) \propto f(r)g(\beta_1)L(r, \beta_1) \geq f(r)(p^E)^{\sum_{j=1}^k n_j}. \quad (\text{A.23})$$

The integration of this lower bound in (A.23) with respect to β_1 is not finite because p^E converges to one as β_1 approaches infinity. Similarly, $\pi_{\text{hyp.post}}(r, \beta_1 \mid \mathbf{y})$ is improper if all the extreme groups have only failures ($y_j = 0$ for all j).

Appendix B

Details on conditional distributions for the Gibbs sampler, profile likelihood, and sensitivity analysis in Chapter 2

B.1 Conditional distributions of $\mathbf{X}(\cdot)$

We define a combined light curve $z(\cdot)$ as follows. The observed magnitude at time t_j^Δ is denoted by $z(t_j^\Delta)$, which is either $x(t_j)$ or $y(t_j - \Delta) - \mathbf{w}_m^\top(t_j - \Delta)\boldsymbol{\beta}$ depending on whether t_j^Δ is one of \mathbf{t} or one of $\mathbf{t} - \Delta$. The observed measurement error is denoted by $\xi(t_j^\Delta)$, which is either $\delta(t_j)$ for $x(t_j)$ or $\eta(t_j)$ for $y(t_j - \Delta) - \mathbf{w}_m^\top(t_j - \Delta)\boldsymbol{\beta}$. We also define $z'(t_j^\Delta)$ as the centered observed magnitude at time t_j^Δ , which is either $x(t_j) - \mu$ or $y(t_j - \Delta) - \mathbf{w}_m^\top(t_j - \Delta)\boldsymbol{\beta} - \mu$ for $j = 1, 2, \dots, 2n$. We introduce the centered latent magnitudes $\mathbf{X}'(\mathbf{t}^\Delta) = \mathbf{X}(\mathbf{t}^\Delta) - \mu$ for notational simplicity. Also, let “ $< t_j^\Delta$ ” denote the set $\{t_i^\Delta : i = 1, 2, \dots, j - 1\}$ and “ $> t_j^\Delta$ ” denote $\{t_i^\Delta : i = j + 1, j + 2, \dots, 2n\}$. We sample $p(\mathbf{X}'(\mathbf{t}^\Delta) \mid \Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau, \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$ by sequentially sampling from its complete conditional distributions, and return $\mathbf{X}'(\mathbf{t}^\Delta)$ to the non-centered latent magnitudes, i.e., $\mathbf{X}(\mathbf{t}^\Delta) = \mathbf{X}'(\mathbf{t}^\Delta) + \mu$ at the end of sampling. To

save space, we suppress conditioning on $\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau, \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t})$. The complete conditional distributions are given by

$$X'(t_1^\Delta) \mid \mathbf{X}'(> t_1^\Delta) \sim \text{N} \left[(1 - B_1^\Delta) z'(t_1^\Delta) + B_1^\Delta a_2 X'(t_2^\Delta), (1 - B_1^\Delta) \xi^2(t_1^\Delta) \right], \quad (\text{B.1})$$

where $B_1^\Delta = \xi^2(t_1^\Delta) / [\xi^2(t_1^\Delta) + \tau\sigma^2(1 - a_2^2)/2]$. For $j = 2, 3, \dots, 2n - 1$,

$$\begin{aligned} & X'(t_j^\Delta) \mid \mathbf{X}'(< t_j^\Delta), \mathbf{X}'(> t_j^\Delta) \\ & \sim \text{N} \left[(1 - B_j^\Delta) z'(t_j^\Delta) + B_j^\Delta \left((1 - B_j) \frac{X'(t_{j+1}^\Delta)}{a_{j+1}} + B_j a_j X'(t_{j-1}^\Delta) \right), (1 - B_j^\Delta) \xi^2(t_j^\Delta) \right], \end{aligned} \quad (\text{B.2})$$

where $B_j^\Delta = \xi^2(t_j^\Delta) / \left[\xi^2(t_j^\Delta) + \frac{\tau\sigma^2}{2} \frac{(1-a_j^2)(1-a_{j+1}^2)}{1-a_j^2 a_{j+1}^2} \right]$ and $B_j = \frac{1-a_{j+1}^2}{1-a_j^2 a_{j+1}^2}$. Lastly,

$$X'(t_{2n}^\Delta) \mid \mathbf{X}'(< t_{2n}^\Delta) \sim \text{N} \left[(1 - B_{2n}^\Delta) z'(t_{2n}^\Delta) + B_{2n}^\Delta a_{2n} X'(t_{2n-1}^\Delta), (1 - B_{2n}^\Delta) \xi^2(t_{2n}^\Delta) \right], \quad (\text{B.3})$$

where $B_{2n}^\Delta = \xi^2(t_{2n}^\Delta) / [\xi^2(t_{2n}^\Delta) + \tau\sigma^2(1 - a_{2n}^2)/2]$ and $a_j = \exp(-(t_j^\Delta - t_{j-1}^\Delta)/\tau)$.

B.2 The likelihood function of parameters.

We use the same notation for the observed data as is defined in Appendix B.1, i.e., $z'(t_j^\Delta)$ and $\xi(t_j^\Delta)$. Let $D_j = \{z'(t_1^\Delta), z'(t_2^\Delta), \dots, z'(t_j^\Delta)\}$. For $j = 1, 2, \dots, 2n$, the posterior predictive density functions of $z'(t_j^\Delta)$ with $\mathbf{X}(\mathbf{t}^\Delta)$ integrated out are

$$z'(t_1^\Delta) \sim \text{N} \left[0, \xi(t_1^\Delta)^2 + \tau\sigma^2/2 \right], \quad (\text{B.4})$$

$$z'(t_j^\Delta) \mid D_{j-1} \sim \text{N} \left[a_j \mu_{j-1}, \xi(t_j^\Delta)^2 + a_j^2 \Omega_{j-1} + \tau\sigma^2(1 - a_j^2)/2 \right], \quad (\text{B.5})$$

where $\mu_1 = (1 - B_1) z'(t_1^\Delta)$, $\mu_j = (1 - B_j) z'(t_j^\Delta) + B_j a_j \mu_{j-1}$, $\Omega_j = (1 - B_j) \xi(t_j^\Delta)^2$, $B_1 = \xi(t_1^\Delta)^2 / [\xi(t_1^\Delta)^2 + \tau\sigma^2/2]$, $B_j = \xi(t_j^\Delta)^2 / [\xi(t_j^\Delta)^2 + a_j^2 \Omega_{j-1} + \tau\sigma^2(1 - a_j^2)/2]$. The likelihood function of $(\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau)$ is the product of the Gaussian densities as follows.

$$L(\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau) \propto p(z'(t_1^\Delta)) \times \prod_{j=2}^{2n} p(z'(t_j^\Delta) \mid D_{j-1}). \quad (\text{B.6})$$

Given the values of $(\boldsymbol{\beta}, \mu, \sigma^2, \tau)$, $L(\Delta, \boldsymbol{\beta}, \mu, \sigma^2, \tau)$ is proportional to the marginalized conditional posterior density $p(\Delta \mid \boldsymbol{\beta}, \mu, \sigma^2, \tau, \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$ for $\Delta \in [u_1, u_2]$ and is zero elsewhere.

B.3 Metropolis-Hastings within Gibbs sampler

We specify the steps of the DA MHwG sampler, (2.14)-(2.17), with $\mathbf{x}(\mathbf{t})$ and $\mathbf{y}(\mathbf{t})$ suppressed in the condition. We sample $\boldsymbol{\beta}$ from its Gaussian conditional posterior distribution as follows.

With a n by n diagonal matrix V whose diagonal elements are $\boldsymbol{\eta}^2(\mathbf{t})$,

$$\boldsymbol{\beta} \mid \mu, \sigma, \tau, \mathbf{X}(\mathbf{t}^\Delta), \Delta \sim N_{m+1} \left[J^{-1} \mathbf{W}_m(\mathbf{t} - \Delta)^\top V^{-1} \mathbf{u}, J^{-1} \right], \quad (\text{B.7})$$

where $J \equiv \mathbf{W}_m^\top(\mathbf{t} - \Delta) V^{-1} \mathbf{W}_m(\mathbf{t} - \Delta) + 10^{-5} I_{m+1}$ and $\mathbf{u} \equiv \mathbf{y}(\mathbf{t}) - \mathbf{X}(\mathbf{t} - \Delta)$.

We sample μ from a truncated Gaussian conditional posterior distribution whose support is $[-30, 30]$; with $a_j = \exp(-(t_j^\Delta - t_{j-1}^\Delta)/\tau)$,

$$\mu \mid \sigma^2, \tau, \mathbf{X}(\mathbf{t}^\Delta), \Delta, \boldsymbol{\beta} \sim \quad (\text{B.8})$$

$$N \left[\frac{X(t_1^\Delta) + \sum_{j=2}^{2n} \frac{X(t_j^\Delta) - a_j X(t_{j-1}^\Delta)}{1 + a_j}}{1 + \sum_{j=2}^{2n} \frac{1 - a_j}{1 + a_j}}, \frac{\tau \sigma^2 / 2}{1 + \sum_{j=2}^{2n} \frac{1 - a_j}{1 + a_j}} \right].$$

The parameter σ^2 has an inverse-Gamma conditional posterior distribution; with its prior distribution $p(\sigma^2) \propto \exp(-b_\sigma/\sigma^2)/(\sigma^2)^2 \cdot I_{\{\sigma^2 > 0\}}$,

$$\sigma^2 \mid \mu, \tau, \mathbf{X}(\mathbf{t}^\Delta), \Delta, \boldsymbol{\beta} \sim \quad (\text{B.9})$$

$$\text{IG} \left(n + 1, b_\sigma + \frac{(X(t_1^\Delta) - \mu)^2}{\tau} + \sum_{j=2}^{2n} \frac{[(X(t_j^\Delta) - \mu) - a_j(X(t_{j-1}^\Delta) - \mu)]^2}{\tau(1 - a_j^2)} \right).$$

The conditional posterior density function of τ is known up to a normalizing constant; with its prior distribution $p(\tau) \propto \exp(-1/\tau)/\tau^2 \cdot I_{\{\tau > 0\}}$,

$$p(\tau \mid \mu, \sigma^2, \mathbf{X}(\mathbf{t}^\Delta), \Delta, \boldsymbol{\beta}) \propto \quad (\text{B.10})$$

$$\frac{\exp \left(-\frac{1}{\tau} - \frac{(X(t_1^\Delta) - \mu)^2}{\tau \sigma^2} - \sum_{j=2}^{2n} \frac{[(X(t_j^\Delta) - \mu) - a_j(X(t_{j-1}^\Delta) - \mu)]^2}{\tau \sigma^2 (1 - a_j^2)} \right)}{\tau^{n+2} \cdot \prod_{j=2}^{2n} (1 - a_j^2)^{1/2}} \cdot I_{\{\tau > 0\}}.$$

To sample τ from (B.10), we use a M-H step with a Gaussian proposal density $N[\log(\tau), \phi^2]$ on a logarithmic scale where ϕ is a proposal scale tuned to produce reasonable acceptance rate.

To implement the ASIS, we need a conditional posterior distribution for $\boldsymbol{\beta}$ given $\mathbf{K}(t^\Delta)$ used in (2.26). Let $K'(t^\Delta) \equiv K(t^\Delta) - \mu$, B be a $2n$ by $(m + 1)$ matrix whose j th row is $(\mathbf{w}_m(t_j^\Delta) - a_j \times \mathbf{w}_m(t_{j-1}^\Delta))^\top$, L be a $2n$ by $2n$ diagonal matrix whose j th diagonal element is $\tau\sigma^2(1 - a_j^2)/2$, \mathbf{b} be a $2n$ by 1 vector whose j th element is $K'(t_j^\Delta) - a_j K'(t_{j-1}^\Delta)$, and finally $A \equiv B^\top L^{-1} B + 10^{-5} I(m + 1)$. Then,

$$\boldsymbol{\beta} \mid \mu, \sigma^2, \tau, K'(t^\Delta), \Delta, \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}) \sim N_{m+1} [A^{-1} B^\top L^{-1} \mathbf{b}, A^{-1}]. \quad (\text{B.11})$$

B.4 Profile likelihood approximately proportional to the marginal posterior

We show that $L_{\text{prof}}(\Delta)$ is approximately proportional to $p(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$. Let $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$. Then,

$$p(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t})) = \int L(\Delta, \boldsymbol{\nu}) p(\Delta, \boldsymbol{\nu}) d\boldsymbol{\nu} = k \int L(\Delta, \boldsymbol{\nu}) p(\boldsymbol{\nu} \mid \Delta) d\boldsymbol{\nu}, \quad (\text{B.12})$$

where k is a normalizing constant of the uniform prior distribution for Δ . We put a Jeffreys' prior on $\boldsymbol{\nu}$ given Δ , i.e., $p(\boldsymbol{\nu} \mid \Delta) \propto |I_\Delta(\boldsymbol{\nu})|^{0.5} d\boldsymbol{\nu}$, where $I_\Delta(\boldsymbol{\nu})$ is Fisher information defined as $-E[\partial^2 \log(L(\Delta, \boldsymbol{\nu})) / \partial \boldsymbol{\nu} \boldsymbol{\nu}^\top]$. The resulting $p(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t}))$ is a Jeffreys-integrated marginal likelihood with its uniform prior (Berger et al., 1999). If we can approximate $l(\Delta, \boldsymbol{\nu}) \equiv \log(L(\Delta, \boldsymbol{\nu}))$ with respect to $\boldsymbol{\nu}$ by a second-order Taylor's series, e.g., under standard asymptotic arguments, then

$$l(\Delta, \boldsymbol{\nu}) \approx l(\Delta, \hat{\boldsymbol{\nu}}_\Delta) - (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)^\top [-l''(\Delta, \hat{\boldsymbol{\nu}}_\Delta)] (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta) / 2, \quad (\text{B.13})$$

where $\hat{\boldsymbol{\nu}}_\Delta = \arg \max_{\boldsymbol{\nu}} l(\Delta, \boldsymbol{\nu})$, and $l''(\Delta, \hat{\boldsymbol{\nu}}_\Delta) \equiv \partial^2 l(\Delta, \boldsymbol{\nu}) / \partial \boldsymbol{\nu} \boldsymbol{\nu}^\top |_{\boldsymbol{\nu}=\hat{\boldsymbol{\nu}}_\Delta}$, which results in

$$L(\Delta, \boldsymbol{\nu}) \approx \exp \left(l(\Delta, \hat{\boldsymbol{\nu}}_\Delta) - (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)^\top [-l''(\Delta, \hat{\boldsymbol{\nu}}_\Delta)] (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta) / 2 \right). \quad (\text{B.14})$$

Using this, we approximate the marginal posterior density function of Δ by

$$\begin{aligned} p(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t})) &\approx k \times L(\Delta, \hat{\boldsymbol{\nu}}_\Delta) \\ &\times \int \exp \left(- (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta)^\top [-l''(\Delta, \hat{\boldsymbol{\nu}}_\Delta)] (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}}_\Delta) / 2 \right) |I_\Delta(\boldsymbol{\nu})|^{0.5} d\boldsymbol{\nu}. \end{aligned} \quad (\text{B.15})$$

If we replace the Fisher information in (B.15), i.e., $I_{\Delta}(\boldsymbol{\nu})$, with the observed information, $-l''_{\Delta}(\hat{\boldsymbol{\nu}}_{\Delta})$, under standard asymptotic arguments, then the integral in (B.15) leads to $(2\pi)^2$ because the integrand becomes a multivariate Gaussian density up to $(2\pi)^{-2}$. Finally,

$$p(\Delta \mid \mathbf{x}(\mathbf{t}), \mathbf{y}(\mathbf{t})) \approx k \times (2\pi)^2 \times L(\Delta, \hat{\boldsymbol{\nu}}_{\Delta}) = k \times (2\pi)^2 \times L_{\text{prof}}(\Delta) \propto L_{\text{prof}}(\Delta). \quad (\text{B.16})$$

B.5 Sensitivity analyses

To see the influence of prior distributions of τ and σ^2 on the posterior distribution of Δ , we conduct sensitivity analyses, varying the scale and shape parameters of their inverse-Gamma (IG) prior distributions.

As an example, we generate 80 observations based on $(\Delta, \beta_0, \mu, \sigma^2, \tau) = (50, 2, 0, 0.03^2, 100)$. The median observation cadence is 3 days and the measurement errors have a constant standard deviation of 0.005 magnitude.

When it comes to fitting the Bayesian model, we assume for simplicity that $\Delta \sim \text{Unif}[0, 100]$ a priori. We run three Markov chains, each of which has 10,000 iterations after 10,000 burn-ins. The initial values of Δ for the three chains are 20, 50, and 80, and those of (μ, σ, τ) are $(0, 0.01, 200)$ the same for every chain. The initial value of β_0 is the average of $\mathbf{y}(\mathbf{t})$ minus that of $\mathbf{x}(\mathbf{t})$. The initial proposal scales, ψ and ϕ , are 10 and 3 days, respectively.

B.5.1 Sensitivity analysis of the prior distribution of τ

We look into the sensitivity of the posterior distribution of Δ to the shape parameter of the IG prior distribution of τ . We denote the shape parameter by a_{τ} and fix the scale parameter at 1. A reasonably small value of the scale parameter does not make any differences in the resultant posterior distribution of τ , not to mention that of Δ , because the conditional posterior density of τ in Appendix B.10 has exponential functions of τ in the exponent, which dominates the scale parameter. We fix the $\text{IG}(1, b_{\sigma})$ prior distribution for σ^2 , where $b_{\sigma} = 8 \times 10^{-6} \text{ mag}^2$ per day as explained in Appendix B.5.2.

Figure B.1 shows the result of sensitivity analysis varying a_τ , the half of the degree of freedom in the corresponding inverse- χ^2 distribution; 0.1, 1, 10, 40, and 80 from the first column. Each column shows the posterior distribution of Δ (first row), that of $\log(\tau)$ (second row), and a scatter plot of posterior samples of $\log(\sigma)$ over those of $\log(\tau)$ (third row) obtained under each shape parameter. The dashed red lines indicate the generative true values.

The first four posterior distributions of Δ catch the generative value of Δ near the mode. However, when we assume too much information in the prior distribution by setting a_τ to 80 ($= n$), the posterior distribution of Δ in the last column becomes flat. A large a_τ concentrates the prior density on O-U processes with mean-reversion timescales τ much shorter than the observational cadence. A large value of a_τ moves the prior mode, $1/(1+a_\tau)$, close to zero and a large degree of freedom ($2 * a_\tau$) for the prior distribution strongly influences the posterior of τ . Hence, the underlying light curves $\mathbf{X}(t^\Delta)$ governed by these O-U processes with small

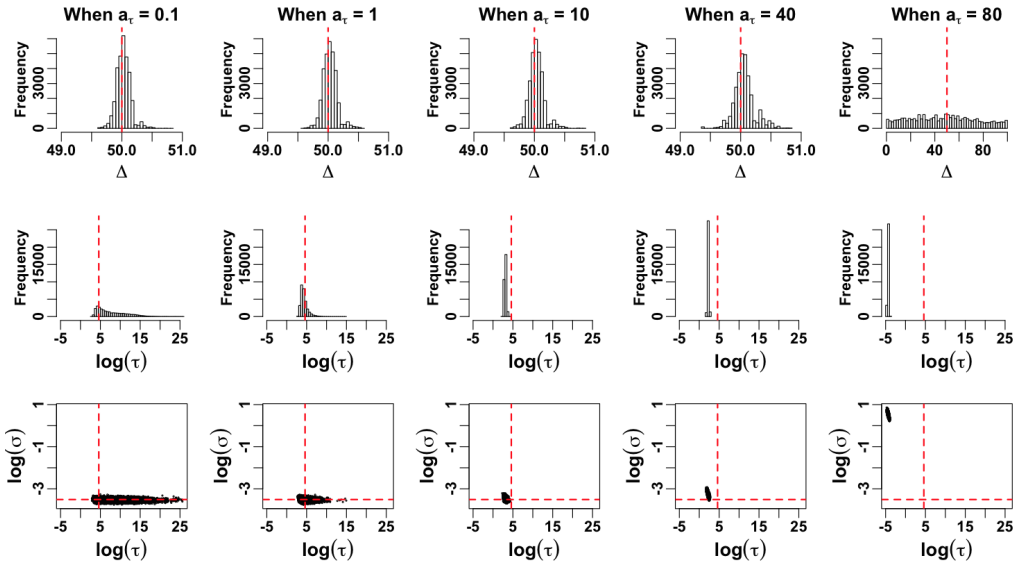


Figure B.1: Each column shows posterior distribution of Δ (first row), that of $\log(\tau)$ (second row), and a scatter plot of $\log(\sigma)$ over $\log(\tau)$ (third row) obtained under a certain a_τ equal to 0.1, 1, 10, 40, and 80 from the first column. The true values of $(\Delta, \log(\sigma), \log(\tau))$ are (50, -3.5, 4.6) represented by the dashed red lines on each plot. The posterior distribution of the time delay is robust to the shape parameter (a_τ) as long as it is reasonably small.

τ will effectively appear as white noise time series. The model will then be ineffective at constraining the time delay because this requires matching serially correlated fluctuation patterns in the light curves. The second row in Figure B.1 shows that as a_τ increases, the mode of the posterior distribution of $\log(\tau)$ gets smaller with a shorter right tail, getting farther away from the generative true value of $\log(\tau) = 4.6$. When the mode of $\log(\tau)$ reaches -5 ($\tau = \exp(-5) = 0.007 \ll 3$ -day observation cadence), the posterior distributions of Δ becomes flat.

B.5.2 Sensitivity analysis of the prior distribution of σ^2

We check the sensitivity of the posterior distribution of Δ to the scale parameter b_σ of the $\text{IG}(1, b_\sigma)$ prior distribution of σ^2 . The effect of the unit shape parameter is negligible because the resultant shape parameter of the IG conditional posterior distribution of σ^2 in (B.9) is $n + 1$ in which n plays a dominant role in controlling the right tail behavior. We fix the $\text{IG}(1, 1)$ prior distribution for τ as explained in the previous section.

We display the result of the sensitivity analysis in Figure B.2, where the values of b_σ are increasing from 0.001 to 10 from the first column. As the soft lower bound ($= b_\sigma/2$) increases from the left, the posterior distribution of the time delay becomes flatter. This is because the true value of σ^2 ($= 0.03^2$) is less than the soft lower bound. For example, when $b_\sigma = 10$ in the last column, the $\text{IG}(1, 10)$ prior distribution of σ^2 exponentially cuts off the probability density in the region to left of the mode, 5 mag² per day, which includes the true value of σ^2 ($= 0.03^2$). Because the true σ^2 is much smaller than the soft lower bound, it is hard for the posterior distribution of σ^2 to move towards the true σ^2 by overcoming the exponential cut-off as the sample size is small. Also, due to the negative correlation between the posterior samples of τ and σ^2 as shown in the scatter plots, the larger the posterior sample of σ^2 is, the smaller the posterior sample of τ is. As the posterior samples of τ become smaller than the observational cadence, the posterior latent light curve $\mathbf{X}(t^\Delta)$ effectively becomes a white noise sequence. In this case, it is difficult to constrain the time delay.

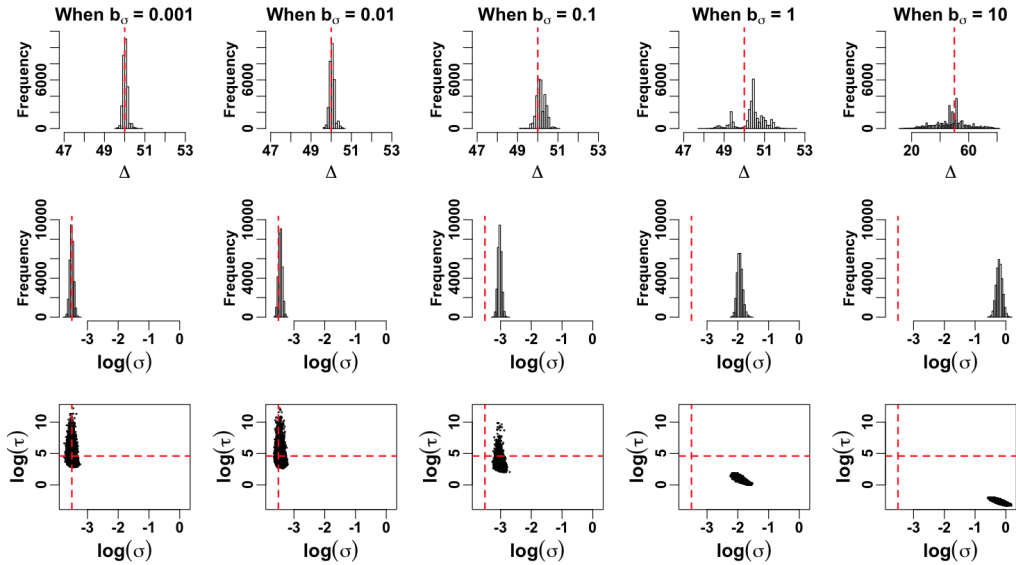


Figure B.2: Each column shows posterior distribution of Δ (first row), that of $\log(\sigma)$ (second row), and a scatter plot of $\log(\tau)$ over $\log(\sigma)$ (third row) obtained under a certain scale parameter b_σ equal to 0.001, 0.01, 0.1, 1, and 10 from the first column. The true values of $(\Delta, \log(\sigma), \log(\tau))$ are $(50, -3.5, 4.6)$ represented by the dashed red lines on each plot. The posterior distributions of parameters recover the true values as the scale parameter (soft lower bound) decreases.

The second row of Figure B.2 shows that as the soft lower bound decreases from the right, the posterior distribution of $\log(\sigma)$ moves towards the true value of $\log(\sigma) = -3.5$ and starts containing it near the mode from the second column ($b_\sigma = 0.01$). The posterior distributions obtained under a value of b_σ smaller than 0.001 do not make noticeable differences, though not shown here. The small soft lower bound also helps the posterior samples of the other parameters cover their true values near the modes. For reference, although it may depend on data, the results became insensitive to the scale b_σ as the number of observations was more than 400.

Bibliography

- Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.
- Albert, J. H. (1988). Computational methods using a bayesian hierarchical generalized linear model. *Journal of the American Statistical Association*, 83(404):1037–1044.
- Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725.
- Athreya, K. B. and Roy, V. (2014). Monte carlo methods for improper target distributions. *Electronic Journal of Statistics*, 8(2):2664–2692.
- Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.
- Behrens, G., Friel, N., and Hurn, M. (2012). Tuning tempered transitions. *Statistics and Computing*, 22(1):65–78.
- Berger, J. O., Liseo, B., Wolpert, R. L., et al. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, 14(1):1–28.
- Berk, D. E. V., Wilhite, B. C., Kron, R. G., Anderson, S. F., Brunner, R. J., Hall, P. B., Ivezić, Ž., Richards, G. T., Schneider, D. P., York, D. G., et al. (2004). The ensemble photometric variability of 25,000 quasars in the sloan digital sky survey. *The Astrophysical Journal*, 601(2):692.
- Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *The*

Annals of Mathematical Statistics, pages 105–110.

Blandford, R. and Narayan, R. (1992). Cosmological applications of gravitational lensing. *Annual Review of Astronomy and Astrophysics*, 30:311–358.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Chang, K. and Refsdal, S. (1979). Flux variations of qso 0957+ 561 a, b and image splitting by stars near the light path. *Nature*, 282:561–564.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Christiansen, C. and Morris, C. (1996). Fitting and checking a two-level poisson model: Modeling patient mortality rates in heart transplant patients. In Berry, D. and Stangl, D., editors, *Bayesian Biostatistics*, pages 467–501. CRC Press.

Christiansen, C. and Morris, C. (1997). Hierarchical poisson regression modeling. *Journal of the American Statistical Association*, 92(438):pp. 618–632.

Courbin, F., Chantry, V., Revaz, Y., Sluse, D., Faure, C., Tewes, M., Eulaers, E., Koleva, M., Asfandiyarov, I., Dye, S., et al. (2013). Cosmograil: the cosmological monitoring of gravitational lenses ix. time delays, lens dynamics and baryonic fraction in he 0435-1223. *Astronomy and Astrophysics*, 536(A53).

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Canadian Journal of Statistics*, 27(3):567–578.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press.

Dean, C. B. (1992). Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457.

- Dobler, G., Fassnacht, C., Treu, T., Marshall, P. J., Liao, K., Hojjati, A., Linder, E., and Rumbaugh, N. (2015). Strong lens time delay challenge. i. experimental design. *The Astrophysical Journal*, 799:168.
- Efron, B. and Morris, C. (1975). Data analysis using stein’s estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319.
- Fassnacht, C., Pearson, T., Readhead, A., Browne, I., Koopmans, L., Myers, S., and Wilkinson, P. (1999). A determination of h_0 with the class gravitational lens b1608+ 656. i. time delay measurements with the vla. *The Astrophysical Journal*, 527(2):498.
- Fischer, P., Bernstein, G., Rhee, G., and Tyson, J. (1997). The mass distribution of the cluster q0957+561 from gravitational lensing. *The Astronomical Journal*, 113(2):521.
- Fohlmeister, J., Kochanek, C. S., Falco, E. E., Wambsganss, J., Oguri, M., and Dai, X. (2013). A two-year time delay for the lensed quasar sdss j1029+ 2623. *The Astrophysical Journal*, 764(2):186.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472.
- Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood. *Interface Foundation of North America*.
- Geyer, C. J. and Thompson, E. A. (1995). Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920.
- Hainline, L. J., Morgan, C. W., Beach, J. N., Kochanek, C., Harris, H. C., Tilleman, T., Fadely, R., Falco, E. E., and Le, T. X. (2012). A new microlensing event in the doubly imaged quasar q 0957+ 561. *The Astrophysical Journal*, 744(2):104.

- Harva, M. and Raychaudhury, S. (2006). Bayesian estimation of time delays between unevenly sampled signals. pages 111–116.
- Hobert, J. P. and Casella, G. (1996). The effect of improper priors on gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436):1461–1473.
- Hojjati, A., Kim, A. G., and Linder, E. V. (2013). Robust strong lensing time delay estimation. *Physical Review D*, 87(12):123512.
- Inada, N., Oguri, M., Morokuma, T., Doi, M., Yasuda, N., Becker, R. H., Richards, G. T., Kochanek, C. S., Kayo, I., Konishi, K., et al. (2006). Sdss j1029+ 2623: A gravitationally lensed quasar with an image separation of 225. *The Astrophysical Journal Letters*, 653(2):L97.
- Jacobsen, M. (1989). Existence and unicity of mles in discrete exponential family distributions. *Scandinavian Journal of Statistics*, pages 335–349.
- Jones, M. and Faddy, M. (2003). A skew extension of the t-distribution, with applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):159–174.
- Kahn, M. J. and Raftery, A. E. (1996). Discharge rates of medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. *Journal of the American Statistical Association*, 91(433):29–41.
- Kass, R. E. and Steffey, D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, 84(407):pp. 717–726.
- Kelly, B. C., Bechtold, J., and Siemiginowska, A. (2009). Are the variations in quasar optical flux driven by thermal fluctuations? *The Astrophysical Journal*, 698(1):895.
- Kelly, J. (2014). *Advances in the Normal-Normal Hierarchical Model*. PhD thesis, Harvard University.

- Kochanek, C., Morgan, N., Falco, E., McLeod, B., Winn, J., Dembicky, J., and Ketzeback, B. (2006). The time delays of gravitational lens he 0435–1223: An early-type galaxy with a rising rotation curve. *The Astrophysical Journal*, 640(1):47.
- Kou, S. C., Zhou, Q., and Wong, W. H. (2006). Discussion paper: Equi-energy sampler with applications in statistical inference and statistical mechanics. *The Annals of Statistics*, 34(4):1581–1619.
- Kozłowski, S. and Kochanek, C. S. (2009). Discovery of 5000 active galactic nuclei behind the magellanic clouds. *The Astrophysical Journal*, 701(1):508.
- Kozłowski, S., Kochanek, C. S., Udalski, A., Soszyński, I., Szymański, M., Kubiak, M., Pietrzyński, G., Szewczyk, O., Ulaczyk, K., Poleski, R., et al. (2010). Quantifying quasar variability as part of a general approach to classifying continuously varying sources. *The Astrophysical Journal*, 708(2):927.
- Kumar, S. R., Stalin, C., and Prabhu, T. (2014). h_0 from 11 well measured time-delay lenses. *Astronomy and Astrophysics*, 580(A38).
- Liao, K., Treu, T., Marshall, P., Fassnacht, C. D., Rumbaugh, N., Dobler, G., Aghamousa, A., Bonvin, V., Courbin, F., Hojjati, A., Jackson, N., Kashyap, V., Rathna Kumar, S., Linder, E., Mandel, K., Meng, X.-L., Meylan, G., Moustakas, L. A., Prabhu, T. P., Romero-Wolf, A., Shafieloo, A., Siemiginowska, A., Stalin, C. S., Tak, H., Tewes, M., and van Dyk, D. (2015). Strong Lens Time Delay Challenge: II. Results of TDC1. *The Astrophysical Journal*, 800:11.
- Linder, E. V. (2011). Lensing time delays and cosmological complementarity. *Phys. Rev. D*, 84:123529.
- Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer Science & Business Media.
- LSST Science Collaboration (2009). Lsst science book, version 2.0. *arXiv*, 912.
- MacLeod, C., Ivezić, Ž., Kochanek, C., Kozłowski, S., Kelly, B., Bullock, E., Kimball, A., Sesar, B., Westman, D., Brooks, K., et al. (2010). Modeling the time variability of sdss

- stripe 82 quasars as a damped random walk. *The Astrophysical Journal*, 721(2):1014.
- Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.
- Morgan, C. W., Hainline, L. J., Chen, B., Tewes, M., Kochanek, C. S., Dai, X., Kozłowski, S., Blackburne, J. A., Mosquera, A. M., Chartas, G., et al. (2012). Further evidence that quasar x-ray emitting regions are compact: X-ray and optical microlensing in the lensed quasar q j0158-4325. *The Astrophysical Journal*, 756(1):52.
- Morris, C. and Tang, R. (2011a). Estimating random effects via adjustment for density maximization. *Statistical Science*, 26(2):pp. 271–287.
- Morris, C. N. and Christiansen, C. L. (1997). Hierarchical poisson regression modeling. *Journal of the American Statistical Association*, 92(438):618–632.
- Morris, C. N. and Lysy, M. (2012). Shrinkage estimation in multilevel normal models. *Statistical Science*, 27(1):115–134.
- Morris, C. N. and Tang, R. (2011b). Estimating random effects via adjustment for density maximization. *Statistical Science*, 26(2):271–287.
- Munoz, J., Falco, E., Kochanek, C., Lehár, J., McLeod, B., Impey, C., Rix, H.-W., and Peng, C. (1998). The castles project. *Astrophysics and Space Science*, 263(1-4):51–54.
- Natarajan, R. and Kass, R. E. (2000). Reference bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95(449):227–237.
- Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6(4):353–366.
- Oscoz, A., Alcalde, D., Serra-Ricart, M., Mediavilla, E., Abajas, C., Barrena, R., Licandro, J., Motta, V., and Munoz, J. (2001). Time delay in qso 0957+ 561 from 1984-1999 optical data. *The Astrophysical Journal*, 552(1):81.

- Oscoz, A., Mediavilla, E., Goicoechea, L. J., Serra-Ricart, M., and Buitrago, J. (1997). Time delay of qso 0957+ 561 and cosmological implications. *The Astrophysical Journal Letters*, 479(2):L89.
- Pelt, J., Hoff, W., Kayser, R., Refsdal, S., and Schramm, T. (1994). Time delay controversy on qso 0957+ 561 not yet decided. *arXiv preprint astro-ph/9401013*.
- Pelt, J., Kayser, R., Refsdal, S., and Schramm, T. (1996). The light curve and the time delay of qso 0957+ 561. *arXiv preprint astro-ph/9501036*.
- R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91.
- Refsdal, S. (1964). The gravitational lens effect. *Monthly Notices of the Royal Astronomical Society*, 128:295–306.
- Robert, C. and Casella, G. (2013). *Monte Carlo Statistical Methods*. Springer Science & Business Media.
- Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, pages 458–475.
- Schneider, P., Ehlers, J., and Falco, E. (1992). *Gravitational Lenses*. Springer.
- Schneider, P., Wambsganss, J., and Kochanek, C. (2006). *Gravitational Lensing: Strong, Weak and Micro*. Springer-Verlag, New York.
- Serra-Ricart, M., Oscoz, A., Sanchís, T., Mediavilla, E., Goicoechea, L. J., Licandro, J., Al-

- calde, D., and Gil-Merino, R. (1999). Bvri photometry of qso 0957+ 561a, b: Observations, new reduction method, and time delay. *The Astrophysical Journal*, 526(1):40.
- Skellam, J. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):257–261.
- Speckman, P. L., Lee, J., and Sun, D. (2009). Existence of the mle and propriety of posteriors for a general multinomial choice model. *Statistica Sinica*, pages 731–748.
- Strawderman, W. E. (1971). Proper bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1):385–388.
- Suyu, S., Auger, M., Hilbert, S., Marshall, P., Tewes, M., Treu, T., Fassnacht, C., Koopmans, L., Sluse, D., Blandford, R., et al. (2013). Two accurate time-delay distances from strong lensing: Implications for cosmology. *The Astrophysical Journal*, 766(2):70.
- Tamura, R. N. and Young, S. S. (1987). A stabilized moment estimator for the beta-binomial distribution. *Biometrics*, pages 813–824.
- Tang, R. (2002). *Fitting and Evaluating Certain Two-Level Hierarchical Models*. PhD thesis, Harvard University.
- Tewes, M., Courbin, F., and Meylan, G. (2013). Cosmograil: the cosmological monitoring of gravitational lenses xi. techniques for time delay measurement in presence of microlensing. *The Astrophysical Journal*, 605(1):58.
- Tierney, L. (1994a). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728.
- Tierney, L. (1994b). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728.
- Uhlenbeck, G. E. and Ornstein, L. S. (1930). On the theory of the brownian motion. *Physical review*, 36(5):823.

- van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1).
- Walsh, D., Carswell, R., and Weymann, R. (1979). 0957+ 561 a, b- twin quasistellar objects or gravitational lens. *Nature*, 279(5712):381–384.
- Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 31(2):144–148.
- Yu, Y. and Meng, X.-L. (2011). To center or not to center: That is not the question—an ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570.
- Zu, Y., Kochanek, C., Kozłowski, S., and Udalski, A. (2013). Is quasar optical variability a damped random walk? *The Astrophysical Journal*, 765(2):106.