ABSTRACT

Poisson Regression Models for Interval Censored Count Data

Sydeaka P. Watson, Ph.D.

Chairpersons: John W. Seaman, Jr., and James D. Stamey

In this dissertation, we develop Bayesian models for interval censored Poisson counts in the presence of zero inflation and missing data. As a motivating example, we consider data arising from a Human Immunodeficiency Virus (HIV) vaccine trial featuring imprecise counts, missing data, and an abundance of values which are either exactly observed to be zero or are left censored. We compare frequentist and Bayesian generalized linear mixed models of the lower limits of the intervals when the data contain no missing values. We then propose a likelihood which models the lower and upper limits of the observed intervals and accomodates zero inflation. Next, we present a simulation study comparing models of the intervals or lower limits to the precise count models. Finally, we apply the model of interval-censored Poisson counts to the HIV data and discuss the conclusions that are drawn from each analysis.

Poisson Regression Models for Interval Censored Count Data

by

Sydeaka P. Watson, B.S., M.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

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

Approved by the Dissertation Committee

_____

John W. Seaman, Jr., Ph.D., Co-Chairperson

_____

James D. Stamey, Ph.D., Co-Chairperson

_____

Dennis Johnston, Ph.D.

_____

Tim Kayworth, Ph.D.

_____

Bette T. Korber, Ph.D.

_____

Jack D. Tubbs, Ph.D.

Accepted by the Graduate School
May 2011

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

Copyright © 2011 by Sydeaka P. Watson
All rights reserved

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to thank Dr. Dean Young and my advisors, Dr. John Seaman and Dr. James Stamey, for giving me the courage and strength to see this part of my academic journey to the end. You revealed greatness in me that I often didn't see in myself. I appreciate your patience, flexibility, wisdom, and kindness.

I thank Dr. Bette Korber for showing me the true meaning of mentorship. I am humbled and honored that you extended the opportunity to become involved in such an exciting research project. You have shaped my career in ways that I never dreamed were possible, and I will always be grateful to you.

I respect and appreciate Robert Gilbert, the first African-American to graduate from Baylor University. Your courage in the face of adversity paved the way for me to harness my potential and use it to change the world. I am proud to honor you through my committment to academic excellence. I will never forget you.

I thank Lindsay Renfro, my most statistically significant friend, for showing me that sometimes the scenic route is the most beautiful. Thank you for daring to be different with me.

I appreciate the support of my best friend and confidant, Wardell Picquet. You listened to my laughter and tears with an objective ear and always responded in love. Thank you for being here.

I thank my mother and brother, Vanessa and Sivad Watson, for loving me because of and in spite of my faults. I am proud to be a product of the love and support you have freely given over the years. This is our success, and I look forward to sharing the rest of this journey with you. I love you.

Finally, I acknowledge the physical manifestations of poetry that I have been blessed to encounter: Baylor University Diverse Verses Poetry Group and the Austin, Dallas, Killeen, New Orleans, and San Antonio poetry communities. Thanks for giving me space to harness my frustrations and turn them into something beautiful. I am a better writer, speaker, lover, and friend because of you. Thank you.

DEDICATION

To my loudest cheerleeder, toughest critic, and softest shoulder:

I love you, Mom.

CHAPTER ONE

Introduction

1.1 Introduction

In this dissertation, we develop Bayesian models for interval censored Poisson counts in the presence of zero inflation and missing data. As a motivating example, we consider data arising from a Human Immunodeficiency Virus (HIV) vaccine trial featuring imprecise counts, missing data, and an abundance of values which are either exactly observed to be zero or are left censored. We compare frequentist and Bayesian generalized linear mixed models of the lower limits of the intervals when the data contain no missing values. We then propose a likelihood which incorporates the lower and upper limits of the observed intervals and accomodates zero inflation. Next, we present a simulation study comparing models of the intervals or lower limits to the precise count models. Finally, we apply the model of interval-censored Poisson counts to the HIV data and discuss the conclusions that are drawn from each analysis.

In Chapters 1 and 2, we examine the frequentist generalized linear mixed model of the lower limits of the HIV immune response counts as presented in Santra et al. (2010). This analysis did not include cases with incomplete observations. We graphically demonstrate the potential selection bias introduced when analyzing the data using only the complete cases. We build an analogous Bayesian model of the lower limits and compare the parameter estimates and inference to those obtained via maximum likelihood estimation.

In Chapter 3, we present a model introduced in Pruszynski (2010) for a single inverval censored Poisson count. We then extend this model to a collection of independent and identically distributed intervals around Poisson counts. We further extend

this model to a regression in which the Poisson count in each interval has a rate parameter that depends on a covariate vector and a set of regression coefficients.

In Chapter 4, we conduct a simulation study of the interval censored Poisson regression models. The data generation method is designed to produce data of similar form to the HIV example dataset. Given binary covariate values and known regression coefficients, our algorithm converts them to Poisson rates according to a log linear model, generates Poisson counts, and generates an interval around each count. The generated data have varying maximum interval widths and proportions of left censored observations. $N = 500$ datasets generated in this fashion are simultaneously analyzed according to four Bayesian models characterized by the type of response used: (i) lower limit of the interval censored observation, (ii) precisely measured count, (iii) interval censored count, and (iv) interval censored count in a mixture model to accommodate zero inflation.

In the simulation study, we find that our proposed models of interval censored data give more accurate and precise estimates of the regression coefficients in both fixed and mixed effect models. Both interval censoring models give parameter estimates with bias that more closely resembles the bias observed in the precise count models. Accounting for zero-inflation in the interval censoring model appears to effectively decrease bias. In the latter part of Chapter 4, we revisit the HIV example data and investigate whether modeling the intervals has any effect on inference regarding vaccine effects.

We finish the present chapter with a detailed look at various features of the HIV vaccine studies used to illustrate methods in the rest of the dissertation. We consider the experimental design, method of data collection, nature of interval censoring, and missing T cell data. We graphically demonstrate the potential selection bias introduced when analyzing the data using only the complete cases. We also present results from Barouch et al. (2010), another study comparing the same vaccines under similar

experimental conditions, including two of the same vaccines and one of the proteins studied in Santra et al. (2010). The data in this study were interval censored counts and featured no missing data. A priori, we expect to arrive at the same conclusions regarding vaccine effects in in Santra et al. (2010) as we obtain in Barouch et al. (2010): the mosaic vaccine produces a higher immune response than the CON-S vaccine, and this enhanced effect is more prominent among CD8+ T cell responses than CD4+ T cell responses.

The study in Santra et al. (2010) compares two HIV vaccine strategies addressing genetic diversity: HIV-1 global CON-S envelope sequence (CON-S) (Santra et al., 2008) and polyvalent vaccine antigens (mosaics) (Fischer et al., 2007). Construction of the CON-S vaccine begins with an alignment of available M-group HIV-1 gene sequences and forms a new sequence containing the most prevalent amino acid at each position. Mosaic proteins are assembled using a computational method that creates a vaccine optimized to cover the largest possible number of T cell epitopes for a given population of HIV-1 strains. The appendix includes a detailed description of the vaccine designs.

Fischer et al. (2007) compared mosaic and CON-S vaccines to all known HIV-1 strains and counted the number of epitopes recognized by each vaccine. Their study predicted a higher degree of epitope recognition for the mosaic group. Separate experimental studies of mosaics in mice (Kong et al., 2009) and CON-S in monkeys (Santra et al., 2008) showed that each vaccine produced higher immune responses than control groups vaccinated with a single HIV strain.

Two recent studies were designed to compare the relative number of immune responses elicited when cells harvested from vaccinated animals were exposed to various HIV-1 strains. Specifically, the goal of each study was to determine if animals given the mosaic vaccine have an advantage over those given the CON-S vaccine. The data in Barouch et al. (2010) and Santra et al. (2010) were analyzed using generalized linear mixed models for Poisson counts, controlling for vaccine type (mosaic or CON-S),

protein type (Gag, Env, Pol in Barouch et al. and Gag, Nef in Santra et al.), T cell type (CD4 or CD8), and random animal effect.

The method of data collection used in Santra et al. (2010) yields counts that are imprecisely measured. In an ELISPOT assay (Miyahira et al., 1995), cells from vaccinated subjects are exposed to fragments of a given HIV-1 strain, or peptides. The level of the subject's immune response is measured by the number of the subject's T cells which recognize the stimulus and signal other cells to attack the intruder. Due to cost and time constraints and a limited number of cells available for testing from each animal, it is not possible to consider all relevant peptides. Instead of testing all peptide sequences of length 9 (or *9-mers*) as in the ideal case, Figure 1.1 illustrates an HIV-1 strain (top) which is segmented into smaller peptide sequences of length 15 overlapping by 10 amino acids. Testing this collection of 15-mers may not yield precise counts of the the numbers of peptides producing positive immune responses. It is not clear whether two consecutive positively responding peptides react independently or if the two reactions correspond to a single 9-mer in their overlap. The true count is greater than or equal to a conservative score which counts consecutive positively responding peptides as a single positive. The count is also less than or equal to the maximum possible count which is realized if no positive reactions occured in the overlaps. Thus, the data are interval censored. Santra et al. modeled the minimum number of positively reacting epitopes in a generalized linear mixed Poisson model. As we shall see in Chapter 4, this is generally a conservative approach, and can be improved by modeling the interval censoring.

The bar graphs in Figure 1.2 show the minimum number of positively reacting epitopes per strain per animal for all subjects in the Santra et al. study. Animals whose IDs are marked with an asterisk did not have a sufficient number of cells available for testing whether the responses should be classified as CD4 or CD8 T cells. While the total number of positive responses per strain is known for all subjects, the number

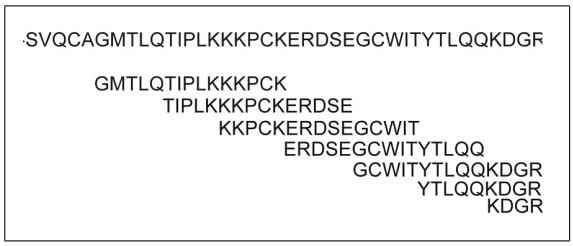

Figure 1.1: Overlapping peptides: This figure illustrates a fragmentation of an HIV-1 strain (top) into peptides 15 amino acids long, overlapping by 10.

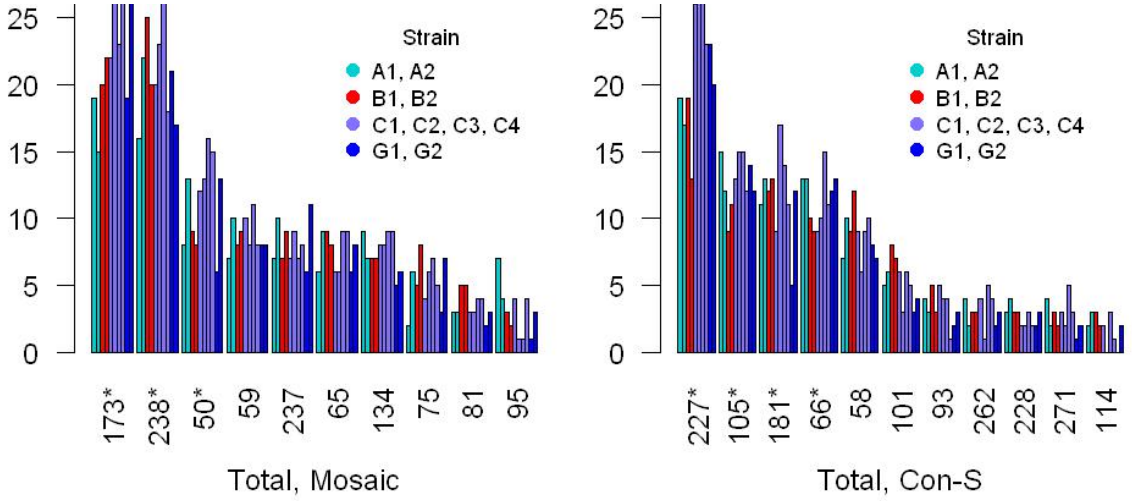

Figure 1.2: Minimum number of positively reacting epitopes for each of ten strains from clades A (A1, A2), B (B1, B2), C (C1, C2, C3, C4), and G (G1, G2) for all subjects, ordered by decreasing median number of positives per animal. Information regarding T cell type is not available for animals whose IDs are marked with an asterisk.

of CD4 (or CD8) responses is missing for seven animals. Preliminary results in both studies suggest that vaccine effect is confounded by T cell type: mosaic outperforms CON-S among CD8 T cells, with the opposite effect observed for CD4 T cells. In the next section, we detail a generalized linear model for the complete cases which controls for T cell effect.

1.2 A Generalized Linear Model for Complete Cases

We now consider a generalized linear model controlling for T cell effect, analyzing a restricted data set which includes the 14 complete cases from the sample of 21 animals. Using only the complete cases, we model the number of immune responses as Poisson counts in a generalized linear mixed model. Let response $Y_{i,t}$ be the minimum number of positive T cell responses observed for subject $i \in \{1, 2, ..., n\}$ and covariate level $t \in \{1, 2, 3, 4\}$. Binary variables are used to describe the three main fixed effects: $x_{2i,t}$ = Vaccine type (=1 for mosaic or 0 for CON-S), $x_{3i,t}$ = T-cell type (=1 for CD8 or 0 for CD4), and $x_{4i,t}$ = region of the HIV-1 genome where the response was elicited (= 1 for Gag or 0 Nef). Combinations of T cell and protein type determine four covariate levels: CD4 and Gag ($t = 1 : \quad x_{3i,1} \equiv 0, \quad x_{2i,1} \equiv 0$), CD4 and Nef ($t = 2 : \quad x_{3i,2} \equiv 0, \quad x_{2i,2} \equiv 1$), CD8 and Gag ($t = 3 : \quad x_{3i,3} \equiv 1, \quad x_{2i,3} \equiv 0$), and CD8 and Nef ($t = 4 : \quad x_{3i,4} \equiv 1, \quad x_{2i,4} \equiv 1$). Two- and three-way interactions ($x_{5i,t} = x_{2i,t} * x_{3i,t}, \quad x_{6i,t} = x_{3i,t} * x_{4i,t}, \quad x_{7i,t} = x_{2i,t} * x_{4i,t}, \quad x_{8i,t} = x_{2i,t} * x_{3i,t} * x_{4i,t}$), a random animal effect ν_i, and an overall intercept term ($x_{1i,t} \equiv 1$) are also included in the model. Let $\mathbf{x}_{i,t} = (x_{1i,t}, x_{2i,t}, x_{3i,t}, x_{4i,t}, x_{5i,t}, x_{6i,t}, x_{7i,t}, x_{8i,t})'$ denote the covariate vector for subject i covariate level t, and let $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)'$ be the corresponding vector of regression coefficients. Our primary interest is in the magnitude and statistical significance of β_2, the coefficient summarizing vaccine effect. The model components are summarized in Table 1.1.

The data model is a log-linear fit of the Poisson rate with random animal effect; that is,

$$Y_{i,t} \sim \text{Poisson}\left[\lambda\left(\mathbf{x}_{i,t}, \nu_i\right)\right], \tag{1.1}$$

where

$$\log\left[\lambda_{i,t}\right] = \log\left[\lambda\left(\mathbf{x}_{i,t}, \nu_i\right)\right] = \mathbf{x}'_{i,t}\boldsymbol{\beta} + \nu_i \tag{1.2}$$

and

$$\nu_i \sim \mathrm{N}(0, \sigma_\nu^2). \tag{1.3}$$

Table 1.1: Binary covariates in equation (1.2)

Coefficients	Covariates	Terms
β_1	x_1	Intercept
β_2	x_2	Vaccine
β_3	x_3	T cell
β_4	x_4	Protein
β_5	$x_2 * x_3$	Vaccine $*$ T cell
β_6	$x_3 * x_4$	T cell $*$ Protein
β_7	$x_2 * x_4$	Vaccine $*$ Protein
β_8	$x_2 * x_3 * x_4$	Vaccine $*$ T cell $*$ Protein

Parameters in the log-linear model outlined in equations (1.1), (1.2), and (1.3) are estimated through maximum likelihood estimation (Table 1.2) with the `glmer` function as implemented in the `lme4` package in `R` (R Development Core Team, 2010). Since the interaction among vaccine, T cell, and protein types is likely to be nonzero (p =0.0174), interpretation of the vaccine coefficient is problematic.

We continue our exploration of the vaccine effect by stratifying the data according to T cell type, fitting either CD4 or CD8 T cell data in reduced models that include

Table 1.2: Parameter estimates in the full model.

Term	Coeff.	Estimate	Std. Err.	Pr(>|z|)	lower.95	upper.95
(Intercept)	β_1	0.365	0.187	0.0515	-0.00239	0.732
Vacc	β_2	-0.148	0.268	0.581	-0.672	0.377
Tcell	β_3	-2.42	0.33	2.07e-13	-3.07	-1.78
Prot	β_4	0.0348	0.132	0.792	-0.224	0.293
Vacc:Tcell	β_5	1.36	0.389	0.000476	0.596	2.12
Vacc:Prot	β_6	0.817	0.181	6.37e-06	0.462	1.17
Tcell:Prot	β_7	1.47	0.374	8.52e-05	0.736	2.2
Vacc:Tcell:Prot	β_8	-1.05	0.442	0.0174	-1.92	-0.185

only vaccine and protein effects. Let response $Y_{4i,t}$ and $Y_{8i,t}$ be the minimum numbers of positive CD4 or CD8 responses observed for subject $i \in \{1, 2, ..., n\}$ and covariate level $t \in \{1, 2, 3, 4\}$, where

$$Y_{4i,t} \sim \text{Poisson}(\lambda_{4i,t}) \tag{1.4}$$

and

$$Y_{8i,t} \sim \text{Poisson}(\lambda_{8i,t}). \tag{1.5}$$

Then the corresponding log-linear models are

$$\log[\lambda_{4i,t}] = x_{1i,t}\beta_{41} + x_{2i,t}\beta_{42} + x_{4i,t}\beta_{43} + x_{7i,t}\beta_{44} + \nu_i \tag{1.6}$$

and

$$\log[\lambda_{8i,t}] = x_{1i,t}\beta_{81} + x_{2i,t}\beta_{82} + x_{4i,t}\beta_{83} + x_{7i,t}\beta_{84} + \nu_i. \tag{1.7}$$

Results of the fits to the separated CD4 and CD8 data are given in Table 1.3. The vaccine and protein interaction is insignificant among CD8 counts (p =0.564), and the vaccine effect is directly interpretable. Subjects treated with the mosaic vaccine have a higher predicted CD8 immune response than those in the CON-S group (p =0.00503), with $e^{1.16} = 3.19$ times as many positive responses expected for mosaics. Interpreting the vaccine effect for CD4 responses is again complicated by the significant interaction between vaccine and protein effects (p =6.36e-06).

We again stratify the data according to Protein type and analyze the resulting subsets. Let response $Y_{4gi,t}$ and $Y_{4ni,t}$ be the minimum numbers of positive CD4 Gag or Nef responses observed for subject $i \in \{1, 2, ..., n\}$ and covariate level $t \in \{1, 2, 3, 4\}$, and let $Y_{8gi,t}$ and $Y_{8ni,t}$ be the analogous counts for CD8 T cell responses. The data models are

$$Y_{4gi,t} \sim \text{Poisson}(\lambda_{4gi,t}), \tag{1.8}$$

$$Y_{4ni,t} \sim \text{Poisson}(\lambda_{4ni,t}), \tag{1.9}$$

$$Y_{8gi,t} \sim \text{Poisson}(\lambda_{8gi,t}), \tag{1.10}$$

and

$$Y_{8ni,t} \sim \text{Poisson}(\lambda_{8ni,t}). \tag{1.11}$$

Then the corresponding log-linear models are

$$\log\left[\lambda_{4gi,t}\right] = x_{1i,t}\beta_{4g1} + x_{2i,t}\beta_{4g2} + \nu_i, \tag{1.12}$$

$$\log\left[\lambda_{4ni,t}\right] = x_{1i,t}\beta_{4n1} + x_{2i,t}\beta_{4n2} + \nu_i, \tag{1.13}$$

$$\log\left[\lambda_{8gi,t}\right] = x_{1i,t}\beta_{8g1} + x_{2i,t}\beta_{8g2} + \nu_i, \tag{1.14}$$

and

$$\log\left[\lambda_{8ni,t}\right] = x_{1i,t}\beta_{8n1} + x_{2i,t}\beta_{8n2} + \nu_i. \tag{1.15}$$

Results are given in Table 1.4. The intercept terms in these models represent the effect of the reference vaccine group (CON-S), or $x_{2i,t} = 0$. The coefficient of $x_{2i,t}$ gives the increase or decrease in effect for the mosaic vaccine relative to the CON-S group. Analyses of the CD4 Gag and Nef counts do not find evidence of significant CON-S effect or difference between mosaic and CON-S response levels. While both CD8 Gag and Nef lower limit analyses identify a CON-S vaccine effect, mosaic and CON-S groups only significantly differ among CD4 Gag responses.

Findings of the lower limit analysis do not support the conclusions of prior

Table 1.3: Parameter estimates in mixed effect CD4 and CD8 models.

Subset	Term	Coeff.	Estimate	Std. Err.	Pr(>|z|)	lower.95	upper.95
CD4	(Intercept)	β_{41}	0.272	0.277	0.327	-0.272	0.816
	Vacc	β_{42}	-0.212	0.395	0.591	-0.986	0.562
	Prot	β_{43}	0.0348	0.132	0.792	-0.224	0.293
	Vacc:Prot	β_{44}	0.817	0.181	6.36e-06	0.462	1.17
CD8	(Intercept)	β_{81}	-2.01	0.348	7.38e-09	-2.69	-1.33
	Vacc	β_{82}	1.16	0.414	0.00503	0.35	1.97
	Prot	β_{83}	1.5	0.352	1.97e-05	0.813	2.19
	Vacc:Prot	β_{84}	-0.234	0.405	0.564	-1.03	0.561

theoretical and experimenal studies of the mosaic and CON-S vaccines. For example, Barouch et al. reported evidence of nonzero CON-S effect and higher response levels for the mosaic vaccine group among CD4 T cells. Various factors may have contributed to the conflicting conclusions. First, the method used to introduce the vaccines into the animals in the Barouch's study tends to elicit more CD8 T cell responses, while more CD4 responses are produced in the Santra's study. The analysis in Santra et al. (2010) also does not take full advantage of the fact that the total number of responses (i.e., the sum of the numbers of CD4 and CD8 responses) is known, resulting in a waste of data and experimental resources.

Another important difference was in the samples. Barouch included 7 subjects in each vaccine group. Santra assigned 11 subjects to each vaccine group, but exclusion of incomplete cases reduced the sample size to 7 subjects per vaccine group. Although the reduced study has the same sample size as Barouch, excluding these cases may have introduced selection bias. The magnitudes of the signals for the excluded subjects (Figure 2.4) are among the highest in the mosaic group and lowest in the CON-S group. As indicated in Figure 1.2, the subjects with missing data had the highest median number of positive reactives per animal. Thus, the model's power to detect differences between the vaccine groups is reduced substantially when these cases are excluded.

Table 1.4: Parameter estimates in mixed effect CD4/Gag model.

Subset	Term	Coeff.	Estimate	Std. Err.	Pr(>|z|)	lower.95	upper.95
CD4/Gag	(Intercept)	$\beta_{4g,1}$	0.192	0.375	0.608	-0.543	0.927
	Vacc	$\beta_{4g,2}$	0.554	0.526	0.293	-0.478	1.59
CD4/Nef	(Intercept)	$\beta_{4n,1}$	0.336	0.197	0.0879	-0.0499	0.722
	Vacc	$\beta_{4n,2}$	-0.115	0.28	0.682	-0.664	0.434
CD8/Gag	(Intercept)	$\beta_{8g,1}$	-0.559	0.241	0.0203	-1.03	-0.0868
	Vacc	$\beta_{8g,2}$	0.919	0.319	0.00398	0.293	1.54
CD8/Nef	(Intercept)	$\beta_{8n,1}$	-2.26	0.483	2.91e-06	-3.2	-1.31
	Vacc	$\beta_{8n,2}$	1.1	0.624	0.0767	-0.119	2.33

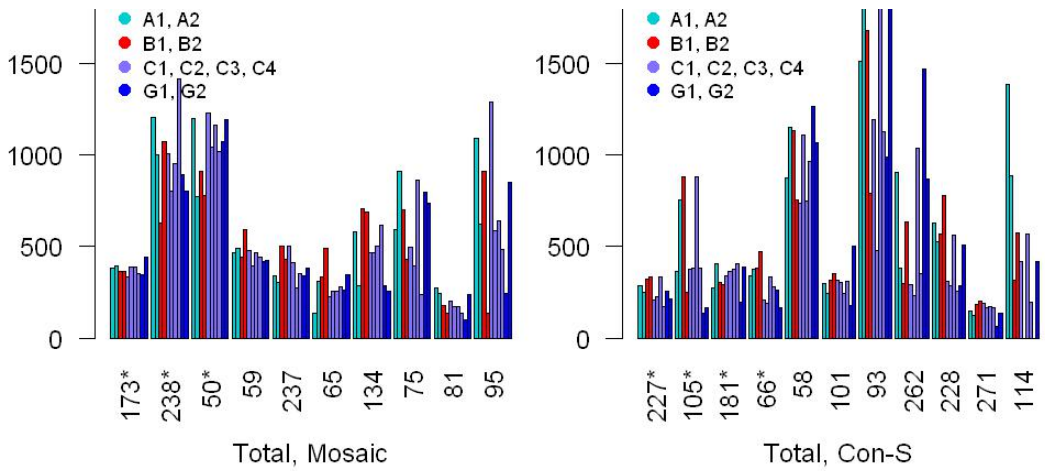

Figure 1.3: The median number of spot forming cells (SFC) per strain per animal.

In Chapter 2, we examine the selection bias potentially introduced by excluding the complete cases. Finally, all immune response counts in Barouch et al. (2010) and Santra et al. (2010) were interval censored. In Chapter 4, we apply interval censored Poisson regression models to the complete cases consider the impact these analyses have on determination of vaccine effects.

CHAPTER TWO

Bayesian Model for the Sum of Two Poissons with Missing Data

In Santra et al. (2010), immune system reaction to HIV-1 exposure may be classified as either a CD4 or CD8 T cell response. Although the total numbers of positive responses are known for all subjects, researchers were not able to count the numbers of each type of T cell response for some subjects. The animals with missing counts of CD4 and CD8 responses were excluded from the non-Bayesian analysis, resulting in a reduction in sample size and the model's power to detect vaccine differences.

In this chapter, the responses are the lower limits of the interval-censored observations. The remaining sections of this chapter will proceed as follows. First, we provide experimental details for two studies of the mosaic and CON-S vaccines. Second, we compare results from the frequentist analysis in Chapter 1 to inference in an analogous Bayesian model. Finally, we discuss the distribution of the magnitudes of the responses, motivating the importance of modeling data for the entire collection of subjects.

2.1 Studies

In the following subsections, we present preliminary results from Barouch et al. (2010) and Santra et al. (2010) and discuss model considerations for the two studies. Experimental details for both studies are provided in subsection 2.1.2.

2.1.1 Barouch et al. (2010) Study

Barouch et al. (2010) compared three vaccines: mosaic vaccine, CON-S vaccine, and another vaccine which was optimized to protect primarily against HIV-1 strains in

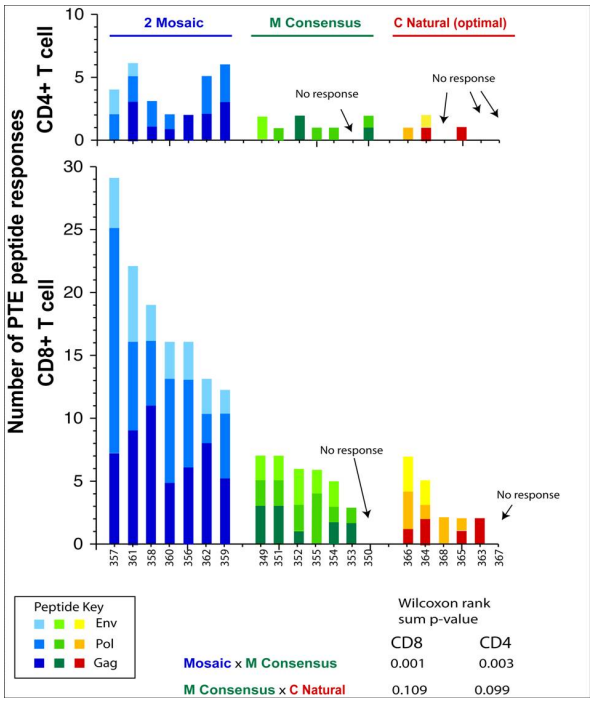

Figure 2.1: Barouch Study Results (Used with permission)

clade C. Primary focus was on eliciting T cell responses from the Gag, Pol, and Env regions of HIV-1. The response, numbers of positively reacting peptides, were interval censored counts. The lower limits of these intervals were modeled as Poisson counts analyzed via a generalized linear mixed model with a random animal effect. The data are summarized in Figure 2.1.

The responses are stratified according to T cell type, and the stacked bars indicate the numbers of immune responses for the Gag, Env, or Pol regions for each animal. Although the mosaic seems to yield more responses for each T cell type, the advantage is enhanced among CD8 groups both overall and within the set of counts corresponding to the Gag protein. The difference between overall counts for mosaic and CON-S groups was detected by the Wilcoxon rank sum test among CD8 ($p = 0.001$) and CD4 ($p = 0.003$) counts.

2.1.2 Santra et al. (2010) Study

The vaccines in Santra et al. (2010) were designed to promote peripheral blood T lymphocyte recognition of HIV-1 Gag and Nef genes. The study investigated the breadth and depth of the immune response (co-authors Watson and Muldoon were the statistical analysts for this study). Breadth measures the number of 9-mers (peptides 9 amino acids long) of HIV-1 that are recognized by the immune system. A vaccine with increased depth corresponds to elicitation of immune responses to more variants (strains) of HIV-1.

To investigate, Santra et al. studied 15-mer peptides (overlapping by 11 amino acids) that completely spanned the HIV-1 Gag and Nef genes. They synthesized 10 sets of HIV-1 Gag and Nef peptides from four clades (2 each from A, B, and G, and 4 from C). These variants were selected based on criteria that considered diversity of the isolates and nation of origin within each clade, availability of the full length of the sequence, how recently these isolates were found relative to the time of the study design (all were sampled after 1999), and prevalence among recently sampled isolates.

Thirty rhesus monkeys were separated into three groups: 12 monkeys received the mosaic vaccine, 12 received the consensus vaccine, and the remaining 6 monkeys were given a placebo. The researchers chose a vaccine delivery mechanism which was appropriate for a vaccine known to produce both CD4+ and CD8+ T lymphocyte responses. After the vaccine regimen, they measured the amount of Anti-Gag and Anti-Nef antibody present at 2 and 4 weeks after the final boost using ELISA (enzyme-linked immunosorbent assay) studies.

Next, they measured the breadth of the cellular immune response using a peptide-stimulated interferon-γ (IFN-γ) enzyme-linked immunospot (ELISPOT) assay. This is a popular technique in which the immune response to HIV-1 exposure is measured *ex-vivo* (i.e., outside of the animal). Cells responding positively to the stimulus release

cytokines which send a signal to other cells to encourage proliferation of T cells specific to the invader. Each well of the ELISPOT assay is coated with antibodies that specifically attract the cytokine of interest in the study. After cells harvested from the vaccinated subjects are added to the wells in the presence of small pools of (or single) peptides from the HIV-1 Gag or Nef genes, they are left to incubate overnight. During this time, the cells release cytokines which are captured by the antibodies coating the wells. When the cells and peptides are removed from the wells, other substances (including another antibody specific for the same cytokineand a dye) are added. The dye marks spots where the cytokine was released by each T cell. Thus, the number of spot-forming cells (SFC) is a measure of the number of T cells which produced an immune response. This number is a measure of the magnitude of the response for a particular peptide stimulus. A positive response is defined as at least 55 SFC per million cells and at least four times the background SFC response (i.e., in the absence of a stimulus). The number of positive responses per strain per animal is a measure of breadth. As discussed in Section 1.2, the counts are interval censored. A discussion of the background measurements and the criteria for positivity is included in the appendix.

One animal in the Mosaic vaccine group (monkey ID #185) did not present an immune response to any of the 10 HIV-1 strains it was exposed to. The biologists deemed this a particularly peculiar outcome since they expected that, even in the absence of a vaccine, the animal's immune system should have produced at least some type of response. Therefore, data for this animal was excluded from the analysis until further review.

For the remaining animals, we modeled the minimum number of immune responses as Poisson counts in a generalized linear mixed model and estimated the regression coefficients using maximum likelihood estimation. This model included Vaccine type (Mosaic or CON-S), Protein type (Gag or Nef), and Vaccine/Protein interaction

term as fixed effects with a random animal effect. The model was plagued with overdispersion, as the estimate of the dispersion parameter in the quasibinomial model was much larger than 1 (McCullagh and Nelder, 1989). The overdispersion vanished when the effect of T cell type (CD4 or CD8) and interactions of T cell type with Vaccine and Protein effects were incoporated into the model. Controlling for T cell effect, however, is only possible for the 14 animals having enough cells for the researchers to make this T cell distinction.

In our approach, we develop an analagous Bayesian model for the complete cases and confirm that the results are similar to those obtained in the frequentist analysis in Santra et al. (2010). The bar graphs in Figure 2.4 illustrate immune response counts per strain per monkey for all animals, with identification numbers for complete cases marked with an asterisk. Mosaic appears to have an advantage over CON-S, but statistical significance of the vaccine effect will be formally investigated in our Bayesian model.

2.2 Models

The data model outlined in equations (1.1), (1.2), and (1.3) was the basis of the frequentist analysis of the HIV data. To estimate the parameters in a Bayesian framework, we employ diffuse normal prior distributions on the regression parameters, assuming *a priori* independence. The random effect is also assumed normally distributed with mean 0. That is,

$$\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, \sigma_\beta^2 \cdot \mathbf{I}) \text{ and} \tag{2.1}$$

$$\sigma_\nu \sim \text{Uniform}(0, B_\nu). \tag{2.2}$$

We follow Gelman (2006) in our use of uniform priors on the scale parameter of the random effect. The initial choices of upper bound B_ν and prior normal variance σ_β^2 are arbitrary. We fit the posterior using Markov chain Monte Carlo methods (MCMC)

implented in `WinBUGS`. Code for this model is in Appendix A.4. A posterior sensitivity analysis (see Appendix A.2) finds Markov chain convergence and minimal effect on posterior quantities of interest for $\sigma_\beta^2 = 100$ and $B_\nu = 1$. Here, we use the `bugs` function (as implemented in the `R2WinBUGS` package in `R`). The analysis involves two Markov chains, a thinning rate of 25 to counter the effects of observed autocorrelation, a burn-in of 25,000 iterations (1,000 iterations retained after thinning) and 125,000 iterations (5,000 iterations retained after thinning).

2.3 Results: Complete Cases

The estimates of the regression coefficients and standard deviation of the random effect term for both saturated models are presented in Table 2.1. The marginal posterior densities of the model parameters for the Bayesian model are displayed in Figure 2.2. As expected, the posterior means and medians closely resemble the maximum likelihood estimates.

The 95% confidence intervals and credible sets from the two models are given for the regression parameters in Table 2.2. Intervals marked with an asterisk contain 0 as an indicator of significance under frequentist criteria. Figure 2.3 graphically illustrates the intervals compared in Table 2.2. The interval estimates are very similar

Table 2.1: Maximum likelihood estimates, posterior means, and posterior medians

	MLE	MLE Std Err	Post Mean	Post Median	Post Std Dev
Intercept	0.365	0.187	0.41	0.412	0.107
Vacc	–0.148	0.268	–0.187	–0.186	0.154
Tcell	–2.425	0.33	–2.454	–2.441	0.342
Prot	0.035	0.132	0.038	0.038	0.148
Vaccine:Tcell	1.358	0.389	1.369	1.361	0.404
Vacc:Prot	0.817	0.181	0.808	0.807	0.203
Tcell:Prot	1.469	0.374	1.492	1.483	0.39
Vacc:Tcell:Prot	–1.051	0.442	–1.052	–1.047	0.464
sigma	0.423		0.346	0.35	0.073

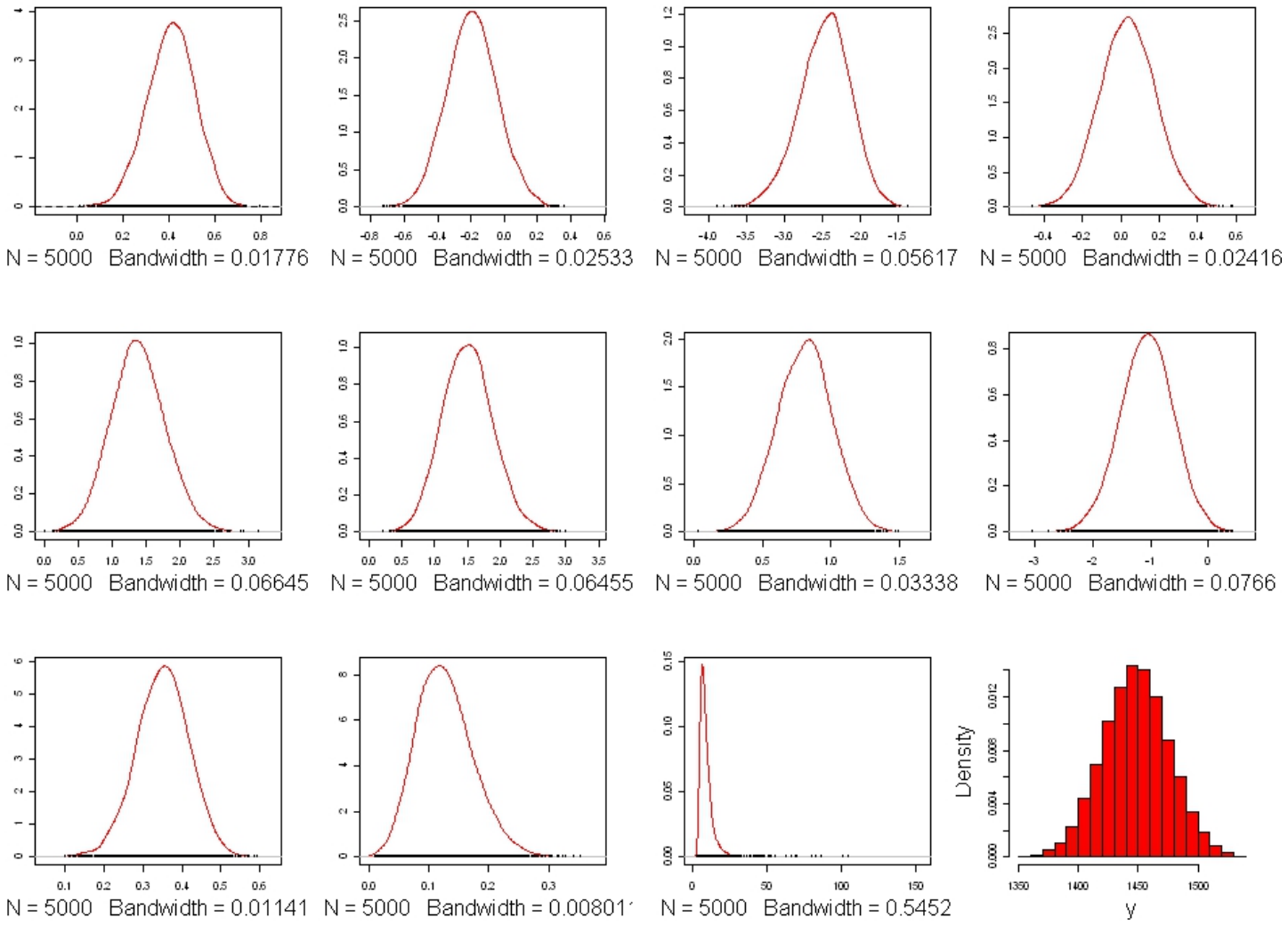

Figure 2.2: Marginal Posteriors of model parameters for the complete cases

Table 2.2: Interval Estimates of model parameters for the complete cases

predictor	95% Conf Interval	95% Cred Set
x_{0i} = Intercept	(–0.01 , 0.73) *	(0.2,0.61)
x_{1i} = Vaccine type	(–0.68 , 0.38) *	(–0.49,0.11) *
x_{2i} = T-cell type	(–3.05 , –1.75)	(–3.17,–1.84)
x_{3i} = Protein type	(–0.22 , 0.29) *	(–0.24,0.32) *
$x_{4i} = x_{1i} \times x_{2i}$	(0.64 , 2.16)	(0.63,2.19)
$x_{5i} = x_{2i} \times x_{3i}$	(0.47 , 1.17)	(0.78,2.28)
$x_{6i} = x_{1i} \times x_{3i}$	(0.77 , 2.23)	(0.42,1.2)
$x_{7i} = x_{1i} \times x_{2i} \times x_{3i}$	(–1.96 , –0.24)	(–1.99,–0.18)
ν_i = Animal		(0.21,0.48)

for all parameters except the intercept. The interval estimate of the intercept in the frequentist model is considerably wider, and its lower limit is approximately equal to zero.

Before assessing the magnitude of the vaccine effect, i.e., β_1, we investigate whether the 3-way interaction (Vaccine/Protein/T cell) is significant. Zero falls outside of the intervals in both models, indicating that the interaction is statistically significant. Both models produced similar point estimates of the standard deviation of the random effect, and the associated credible set is centered near 0.42, the estimate obtained from the frequentist analysis.

2.4 Magnitude of Responses

Next, we compare the magnitudes of the responses for the vaccine groups. The bar graphs in Figure 2.4 summarizes the strength of the immune responses with the median numbers of spot-forming cells (SFC) for all subjects for the 10 strains tested. Excluding the incomplete cases (with IDs marked with asterisks) removes the strongest mosaic responses and weakest CON-S responses. These animals also produced the highest number of responses among mosaics and lowest numbers of responses in the CON-S group. This suggests that excluding these animals may have introduced selection bias.

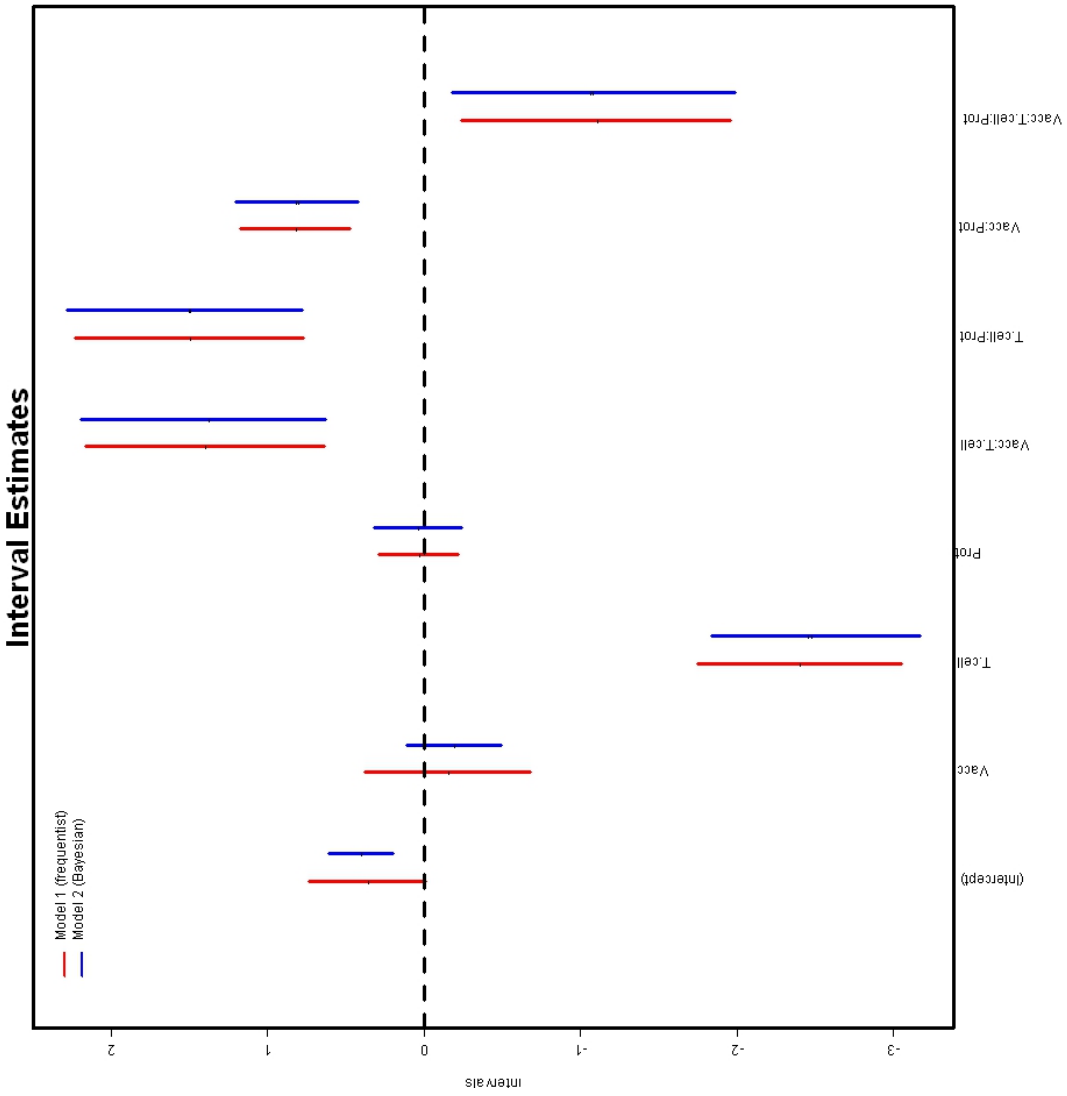

Figure 2.3: Interval Estimates of model parameters for the complete cases

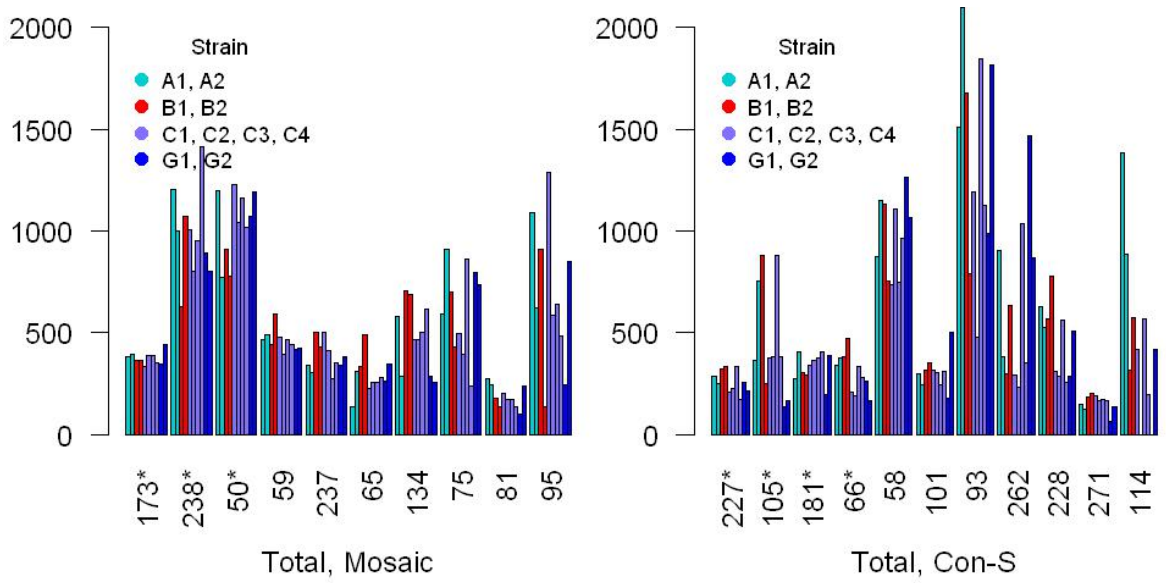

Figure 2.4: Median SFC measurements per strain for all subjects

We proceed with the data at hand, but are continuing to develop a model that will allow us to incorporate the animals with incomplete data into the analysis.

CHAPTER THREE

Interval Censored Poisson Regression

3.1 Introduction

Many studies feature an outcome of interest which is summarized by a single count observation. Popular models for such data include the Poisson, binomial, and negative binomial distributions as appropriate. When it is not possible to precisely measure these counts, we may alternatively observe that the true count lies within an interval with probability 1. These imprecise measurements are known as censored data. A count is *left-censored* if the true value y is known to lie below a given value. For example, we may observe that $0 \leq y \leq c$, where c is known. If the count is known to lie above a given value c, for example, $c \leq y \leq n$ (or $< \infty$, as appropriate), the data are *right-censored*. These types of censoring intervals are bounded on on one side by the extrema of the support of the respective distribution. If the true count is known with probability 1 to lie between two values that do not necessarily coincide with the minimum or maximum of the support (e.g., $c_1 \leq y \leq c_2$), the data are *interval-censored*.

Models of left or right censored Poisson regression models have been extensively studied. For example, see Famoye and Wang (2004) and Terza (1985). Although literature concerning interval censored count data regression models is sparse, researchers have analyzed data of this type in many ways. One method involves dichotomization of left or right censored data into positive and negative categories (Ekanem et al., 1983). Others estimate the true count using one of the conventional imputation methods such as the EM algorithm or multiple imputation (Rubin, 1987). The statistical software package `Stata` offers the `intreg` function (Long and Freese, 2006) which models inter-

val censored responses in a regression context. The technique samples from a normal distribution which is truncated to the censoring limits, and it assumes normal errors.

A simple approach to handling this type of data is to use the lower or upper limit of the interval censored count as a substitute for the true count and proceed with standard univariate count data methods implemented in statistical software. However, this method (and the others discussed above) discards valuable information about the data and reduce the power of the model's ability to detect significant effects. In an HIV-1 vaccine study, for example, analyses of lower limits may conclude a lack of a significant vaccine effect when, in fact, there is one.

Pruszynski (2010) provides a full development of a likelihood for θ given a single interval censored count observation arising from a Poisson(θ) process. Pruszynski introduced the likelihood and posterior distribution under a gamma prior for θ. Although not considered in Pruszynski (2010), the extension of this model to a collection of n independent intervals from the same process is immediate: the likelihood given the n intervals is the product of the individual likelihoods, and the posterior distribution for θ given the collection of intervals is of similar form. This model can be further extended to interval-censored Poisson regression. That is, we can model n interval censoring limits of independent counts arising from Poisson processes with mean parameters $\theta_i \equiv \theta(\mathbf{x}_i, \boldsymbol{\beta})$ that depend on a $q \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ and individual $q \times 1$ covariate vectors $\mathbf{x}_i$, $i = 1, ..., n$. For a one-to-one, differentiable link function $g(\cdot)$, the linear fixed effect model of the data is described by $g(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta}$.

In this chapter, we provide a full development of various interval-censored Poisson likelihoods. First, we present the interval-censored Poisson single observation likelihood introduced in Pruszynski (2010). We discuss an extension of this model to accomodate multiple independent observations from the same underlying Poisson process. Third, we propose a Bayesian model for interval censored fixed effect regression. Feasibility of parameter estimation and proposed restrictions on the amount of censoring are dis-

cussed. We further extend the regression model to include mixed effects. Finally, we revisit the HIV vaccine example and apply the proposed models of the interval-censored counts.

3.2 Likelihood Given a Single Observed Interval

In this section, we consider properties of the interval-censored Poisson likelihood for Poisson rate θ given a single observed interval. In the frequentist paradigm, we derive maximum likelihood estimates and likelihood-based interval estimates. Bayesian properties of the likelihood derived in Pruszynski (2010) are reviewed.

3.2.1 Interval-Censored Poisson Distribution

Let $Y \sim \text{Poisson}(\theta)$ be a count with unknown true value. Suppose that it is interval-censored, so that the true value of Y is observed to lie between positive integers j and k, $j \leq k$, with probability 1. The probability distribution function and cumulative distribution function of the exactly observed data are, respectively

$$f(y|\theta) = P(Y = y|\theta) = \frac{\theta^y e^{-\theta}}{y!}$$

and

$$F(y|\theta) = P(Y \leq y) = \sum_{t=0}^{y} \frac{e^{-\theta}\theta^t}{t!} = \frac{\Gamma(y+1,\theta)}{y!},$$

where $\Gamma(a,\theta) = \int_\theta^\infty t^{a-1}e^{-t}dt$ is the incomplete gamma function. Writing the probability that $Y \in [j,k]$ as a function of θ produces the interval-censored data likelihood function for θ, or

$$L(\theta|j \leq y \leq k) = \sum_{y=j}^{k} \frac{\theta^y e^{-\theta}}{y!}.$$

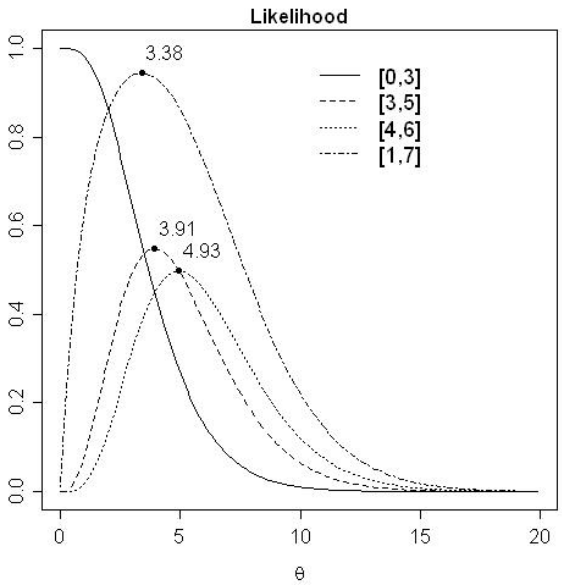

Figure 3.1: Plot of the likelihood given (j, k) pairs (0,3), (3,5), (4,6), and (1,7). Maximum likelihood estimates are shown when they exist.

3.2.2 Regularity of the Interval-Censored Likelihood

Regularity of the likelihood is directly related to the type of interval observed. For example, consider interval censored observations from a Poisson(θ) process. Figure 3.1 below gives four example plots of the likelihood function for (j, k) pairs (0,3), (3,5), (4,6), and (1,7). The black points on the curves, if present, show where the likelihoods are maximized. The graph of the likelihood function is irregular with undefined maximum likelihood estimate for the left-censored case of $j = 0, k = 3$. That is, the form of the likelihood does not feature the approximately quadratic shape desired for maximum likelihood estimation and use of asymptotic results typically needed for construction of 95% confidence intervals. Other curves in Figure 3.1 represent the likelihoods for θ given observed intervals $(3, 5)$, $(4, 6)$, or $(1, 7)$, respectively. Maximum likelihood estimates exist for these regular likelihoods and are shown. For example, given that the unknown count is observed to lie in interval $(3, 5)$, the MLE for theta is 3.91.

3.2.3 Statistical Inference in the Frequentist Paradigm

The probability characterizing the likelihood can be written as a difference of two cumulative probabilities: $P(j \leq Y \leq k) = P(0 \leq Y \leq k) - P(0 \leq Y \leq j-1)$. This suggests the following alternative formulation of the likelihood:

$$\begin{aligned} L(\theta|j \leq y \leq k) &= \sum_{y=j}^{k} \frac{\theta^y e^{-\theta}}{y!} \\ &= \sum_{y=0}^{k} \frac{e^{-\theta}\theta^y}{y!} - \sum_{y=0}^{j-1} \frac{e^{-\theta}\theta^y}{y!} \\ &= \frac{\Gamma(k+1,\theta)}{k!} - \frac{\Gamma(j,\theta)}{(j-1)!}. \end{aligned}$$

This formulation facilitates differentiation of the likelihood for the purposes of maximum likelihood estimation. The derivative of the incomplete gamma function with respect to its second argument is

$$\frac{\partial \Gamma(a,\theta)}{\partial \theta} = -e^{-\theta}\theta^{a-1}.$$

The first derivative of the likelihood with respect to θ is

$$\begin{aligned} \frac{dL(\theta|j \leq y \leq k)}{d\theta} &= -e^{-\theta}\left(\frac{\theta^k}{k!} - \frac{\theta^{j-1}}{(j-1)!}\right) \\ &= -e^{-\theta}\left(\frac{(j-1)!\theta^k - k!\theta^{j-1}}{k!(j-1)!}\right). \end{aligned}$$

Then the MLE for θ given a single observed interval (j,k) is derived as follows. The likelihood equation is

$$-e^{-\hat{\theta}}\left(\frac{(j-1)!\hat{\theta}^k - k!\hat{\theta}^{j-1}}{k!(j-1)!}\right) = 0,$$

which may be written as

$$(j-1)!\hat{\theta}^k - k!\hat{\theta}^{j-1} = 0$$

or

$$(j-1)!\hat{\theta}^k = k!\hat{\theta}^{j-1}.$$

Taking logs, we have

$$\log[(j-1)!] + k \log\left[\hat{\theta}\right] = \log[k!] + (j-1)\log\left[\hat{\theta}\right]$$

or

$$\log\left[\hat{\theta}\right](k-j+1) = \log[k!] - \log[(j-1)!]$$

which implies

$$\log\left[\hat{\theta}\right] = \frac{\log[k!/(j-1)!]}{(k-j+1)}.$$

Solving for $\hat{\theta}$ yields

$$\hat{\theta} = \left[\frac{k!}{(j-1)!}\right]^{1/(k-j+1)}. \tag{3.1}$$

The second derivative of the likelihood is

$$\frac{d^2 L(\theta | j \le y \le k)}{d\theta^2} = e^{-\theta}\ \theta^{j-2}\left[\frac{j-1-\theta}{(j-1)!} + \frac{\theta^{k-j+1}(\theta-k)}{k!}\right].$$

The second derivative evaluated at $\hat{\theta}$ is

$$\begin{aligned}
\left.\frac{d^2 L(\theta | j \le y \le k)}{d\theta^2}\right|_{\theta=\hat{\theta}} &= e^{-\hat{\theta}}\ \hat{\theta}^{j-2}\left[\frac{j-1-\hat{\theta}}{(j-1)!} + \hat{\theta}^{k-j+1}\frac{\hat{\theta}-k}{k!}\right] \\
&= e^{-\hat{\theta}}\ \hat{\theta}^{j-2}\left[\frac{j-1-\hat{\theta}}{(j-1)!} + \left\{\left[\frac{k!}{(j-1)!}\right]^{1/(k-j+1)}\right\}^{k-j+1}\frac{\hat{\theta}-k}{k!}\right] \\
&= e^{-\hat{\theta}}\ \hat{\theta}^{j-2}\left[\frac{j-1-\hat{\theta}}{(j-1)!} + \frac{\hat{\theta}-k}{(j-1)!}\right] \\
&= e^{-\hat{\theta}}\ \hat{\theta}^{j-2}\left[\frac{j-1-\hat{\theta}+\hat{\theta}-k}{(j-1)!}\right] \\
&= e^{-\hat{\theta}}\ \hat{\theta}^{j-2}\left[\frac{j-1-k}{(j-1)!}\right],
\end{aligned}$$

which is negative since $j-1 < k$. Thus, equation (3.1) yields a maximum. This MLE is not defined given a left-censored interval (i.e., for $j = 0$).

3.2.4 Statistical Inference in the Bayesian Paradigm

Pruszynski (2010) proposed a Gamma(α, β) prior distribution for θ. The resulting posterior distribution for θ given the interval censored observations is

$$\begin{aligned}\pi(\theta|j \leq y \leq k) &= \left[\sum_{y=j}^{k} \frac{\theta^y e^{-\theta}}{y!}\right] \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\right] \left[\frac{1}{m(j,k)}\right] \\ &= \left[\frac{1}{m(j,k)}\right] \frac{\beta^\alpha}{\Gamma(\alpha)} \sum_{y=j}^{k} \frac{\theta^{y+\alpha-1}e^{-\theta(\beta+1)}}{y!},\end{aligned}$$

where

$$\begin{aligned}m(j,k) &= \int_0^\infty \left[\sum_{y=j}^{k} \frac{\theta^y e^{-\theta}}{y!}\right] \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\right] d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \sum_{y=j}^{k} \frac{\Gamma(y+\alpha)}{(\beta+1)^{y+\alpha}y!}.\end{aligned}$$

Because this posterior is available in closed form, we may directly sample from this distribution.

3.3 Likelihood Given Multiple Observed Intervals

Let $\{Y_i\}$, $i = 1, ..., n$, be a collection of independent Poisson-distributed random variables with mean θ, such that $Y_i \in [j_i, k_i]$ with probability 1, where j_i and k_i are positive integers and $j_i < k_i$. Then the interval-censored data likelihood function for θ is

$$\begin{aligned}L(\theta|\{j_i \leq y_i \leq k_i\}_{i=1}^n) &= \prod_{i=1}^{n} \sum_{y_i=j_i}^{k_i} \frac{\theta^{y_i}e^{-\theta}}{y_i!} \\ &= e^{-n\theta} \prod_{i=1}^{n} \sum_{y_i=j_i}^{k_i} \frac{\theta^{y_i}}{y_i!}.\end{aligned}$$

3.3.1 Regularity of the Interval-Censored Likelihood

As in the single observation case, regularity is directly related to the nature of censoring observed. The likelihood plots in Figure 3.2 illustrate this relationship. For example, given 2 pairs of left censored intervals (0,3) and (0,4), the likelihood is irregular, with a shape similar to that seen in the case of a single observed left censored interval. Maximum likelihood estimation is not defined for θ given a set of intervals which are all left censored. Another likelihood graph for a pair of intervals including one left censored $(0, 3)$ and one interval censored observation $(3, 5)$ is shown. This curve features a regular likelihood and a unique maximum likelihood estimate.

A third likelihood plot is given for four pairs: three left-censored [(0,3), (0,4)

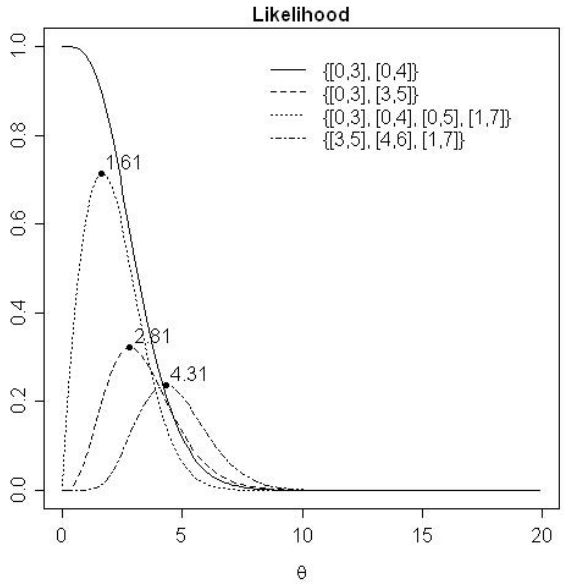

Figure 3.2: Likelihood plots for θ given multiple observations.

(0,5)], and one interval censored, (3,5). The likelihood is still regular, even when there is only one interval which is not left-censored. A fourth graph of three non-left censored observations is also shown, also featuring a regularlikelihood. Although we have not

proven this mathematically, this illustration suggests conditions under which maximum likelihood estimation is possible in this extension of Pruszynski's model.

3.3.2 Statistical Inference in the Frequentist Paradigm

The log-likelihood function is

$$l(\theta|\{j_i \leq y_i \leq k_i\}_{i=1}^n) = -n\theta + \sum_{i=1}^{n} \log \left\{ \sum_{y_i=j_i}^{k_i} \frac{\theta^{y_i}}{y_i!} \right\}.$$

The first derivative of the log-likelihood function with respect to θ is

$$\frac{dl(\theta|\{j_i \leq y_i \leq k_i\}_{i=1}^n)}{d\theta} = -n + \sum_{i=1}^{n} \frac{\sum_{y_i=j_i}^{k_i} \frac{y_i \theta^{y_i-1}}{y_i!}}{\sum_{y_i=j_i}^{k_i} \frac{\theta^{y_i}}{y_i!}}.$$

Differentiating the log likelihood yields a rational polynomial expression. A closed-form solution of the resulting likelihood equation is not apparent, even given an alternative formulation with the incomplete gamma function as in the single observation case. In the absence of a closed form solution, we use numerical optimization procedures to obtain parameter estimates.

3.3.3 Statistical Inference in the Bayesian Paradigm

If we again specify a Gamma(α, β) prior distribution for θ, the posterior distribution for θ given n interval censored observations is

$$\begin{aligned} \pi(\theta|\{j_i \leq y_i \leq k_i\}_{i=1}^n) &\propto \left[\prod_{i=1}^{n} \sum_{y_i=j_i}^{k_i} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \right] \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \right] \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \prod_{i=1}^{n} \sum_{y_i=j_i}^{k_i} \frac{\theta^{y_i+(\alpha-1)/n} e^{-\theta(\beta/n+1)}}{y_i!}. \end{aligned}$$

Unlike the single observation case, the posterior for θ given multiple observations is not available in closed form. We use MCMC methods to study the features of this posterior distribution.

3.4 Fixed Effect Regression Given Multiple Observed Intervals

Let $Y_1, ..., Y_n$ be a collection of n independent count observations from Poisson processes with mean parameters $\theta_i = \theta_i|\mathbf{x}_i, \boldsymbol{\beta}$, where θ_i is a function of a vector of covariates $\mathbf{x}_i$ and regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)'$. Suppose that the counts are interval censored such that $Y_i \in [j_i, k_i]$ with probability 1, where j_i and k_i are positive integers and $j_i < k_i$, $i = 1, ..., n$. Let $g(\cdot)$ be a one-to-one, differentiable link function, such that $g(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta}$. Then $\theta_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$. The probability distribution functions of the exactly observed data are

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = P(Y_i = y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \frac{\theta_i{}^{y_i} e^{-\theta_i}}{y_i!}. \tag{3.2}$$

The interval-censored Poisson (ICP) regression likelihood function for $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}\,|\{j_i \le y_i \le k_i\}_{i=1}^n, \{\mathbf{x}_i\}) = \prod_{i=1}^n \sum_{y_i=j_i}^{k_i} \frac{[g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})]^{\,y_i} e^{-\left[g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]}}{y_i!}. \tag{3.3}$$

Given a multivariate normal prior structure on the regression coefficients, i.e.,

$$\pi(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{q/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}, \tag{3.4}$$

the posterior distribution for $\boldsymbol{\beta}$ given the n interval censored observations is

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\{j_i \le y_i \le k_i\}_{i=1}^n) &\propto \left[\prod_{i=1}^n \sum_{y_i=j_i}^{k_i} \frac{[g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})]^{\,y_i} e^{-\left[g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]}}{y_i!}\right] \\
&\quad \times \left[\frac{1}{(2\pi)^{q/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}\right] \\
&= \frac{1}{(2\pi)^{q/2}|\boldsymbol{\Sigma}|^{1/2}} \prod_{i=1}^n \sum_{y_i=j_i}^{k_i} U_i(y_i)\; V_i,
\end{aligned} \tag{3.5}$$

where

$$U_i(y_i) = \frac{[g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}}{y_i!}$$

and

$$V_i = e^{-g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right)+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}.$$

It may be difficult to accurately estimate the parameters when there is an abundance of left censored observations (i.e., $j = 0$). We propose a modification to the interval-censored regression likelihood which accomodates zero inflation. Each component of the likelihood is a mixture of a point mass at zero and the interval-censored Poisson regression likelihood. For mixture probabilities $p_{01}, ..., p_{0n}$, $0 \leq p_{0i} \leq 1$, $i = 1, ..., n$, and fixed effects regression likelihood contribution $L(\boldsymbol{\beta}|j_i \leq y_i \leq k_i)$ for subject i, the interval-censored Poisson regression likelihood with adjustment for zero-inflation (ICP ZIP) is

$$L_{ZIP}\left(\boldsymbol{\beta}\,|\{j_i \leq y_i \leq k_i\}\,, \{\mathbf{x}_i\}\right) = \prod_{i=1}^{n} \left[p_{0i}\; I(j_i = 0) + (1 - p_{0i})\; L(\boldsymbol{\beta}|j_i \leq y_i \leq k_i)\right], \tag{3.6}$$

where

$$L(\boldsymbol{\beta}|j_i \leq y_i \leq k_i) = \sum_{y_i=j_i}^{k_i} \frac{\left[g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]^{y_i} e^{-\left[g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right]}}{y_i!}.$$

3.5 Mixed Effect Regression Given Multiple Observed Intervals

Let $\mathbf{Y}_1, ..., \mathbf{Y}_n$ be a collection of n independent data vectors, where each vector $\mathbf{Y}_i = (Y_{i,1}, ..., Y_{i,R})'$ is a set of Poisson(θ_i) counts measured on subject i, $i = 1, ..., n$. Suppose that the rate parameters can be represented by the generalized linear mixed model

$$g(\theta_i) \equiv g(\theta_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\nu}_i) = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\nu}_i, \tag{3.7}$$

i.e., as functions of fixed effect covariates $\mathbf{x}_i$ with coefficient vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)'$, and random effect covariates $\mathbf{z}_i$ with coefficient vector $\boldsymbol{\nu}_i = (\nu_{i1}, ..., \nu_{i,v})'$. Further suppose

that the counts are interval censored, such that $Y_{i,r} \in [j_{i,r}, k_{i,r}]$ with probability 1, where $j_{i,r}$ and k_i are positive integers and $j_{i,r} < k_{i,r}$, $r = 1, ..., R$. If $g(\cdot)$ is a one-to-one, differentiable link function, then $\theta_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\nu}_i)$. Hence the mixed effects interval-censored Poisson (ICP) regression likelihood function for $\boldsymbol{\beta}$ is

$$L\left(\boldsymbol{\beta}\left|\left\{\{j_{i,r} \le y_{i,r} \le k_{i,r}; \mathbf{x}_i; \mathbf{z}_i\}_{i=1}^{n}\right\}_{r=1}^{R}\right.\right) = \prod_{i=1}^{n}\prod_{r=1}^{R}\sum_{y_{i,r}=j_{i,r}}^{k_{i,r}} U(y_{i,r})\ V_i, \tag{3.8}$$

where

$$U(y_{i,r}) = \frac{\left[g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\nu}_i\right)\right]^{y_{i,r}}}{y_{i,r}!}$$

and

$$V_i = e^{-\left\{\left[g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\nu}_i\right)\right]+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right\}}.$$

If we again specify a multivariate normal prior as in (3.4), the posterior distribution for $\boldsymbol{\beta}$ given the n interval censored observations is

$$\begin{aligned}\pi(\boldsymbol{\beta}|\{\{j_{i,r} \le y_{i,r} \le k_{i,r}; \mathbf{x}_i; \mathbf{z}_i\}_{i=1}^{n}\}_{r=1}^{R}) &\propto L\left(\boldsymbol{\beta}\left|\{\{j_{i,r} \le y_{i,r} \le k_{i,r}; \mathbf{x}_i; \mathbf{z}_i\}_{i=1}^{n}\}_{r=1}^{R}\right.\right)\pi(\boldsymbol{\beta}) \\ &= \frac{1}{(2\pi)^{q/2}|\boldsymbol{\Sigma}|^{1/2}}\prod_{i=1}^{n}\prod_{r=1}^{R}\sum_{y_{i,r}=j_{i,r}}^{k_{i,r}} U(y_{i,r})\ V_i. \end{aligned} \tag{3.9}$$

Given mixture probabilities $p_{01}, ..., p_{0n}$, we control for zero inflation in the mixed effect regression model according to the following mixture model. The mixed effects interval-censored Poisson regression likelihood with adjustment for zero-inflation (ICP ZIP) is

$$L_{ZIP}\left(\boldsymbol{\beta}\left|\{\{j_{i,r} \le y_{i,r} \le k_{i,r}; \mathbf{x}_i; \mathbf{z}_i\}_{i=1}^{n}\}_{r=1}^{R}\right.\right) = \prod_{i=1}^{n}\prod_{r=1}^{R}\left[p_{0i}\ I(j_{i,r}=0) + (1-p_{0i})\ Q(y_{i,r})\right], \tag{3.10}$$

where

$$Q(y_{i.r}) = \sum_{y_{i,r}=j_{i,r}}^{k_{i,r}} U(y_{i,r})\ V_i.$$

3.6 HIV Example Data with Interval-Censored Poisson Regression

In this section, we revisit the HIV example dataset and apply our interval censoring likelihoods to the complete cases. For all models, the data are analyzed via the MCMC methods implemented in `OpenBUGS`. Code for this implementation is provided in Appendix A.4. We discard the first 15,000 iterations, run the algorithm for 15,000 iterations after the burn-in period, and keep every tenth iteration to address autocorrelation issues. Other convergence diagnostics, including history plots and Gelman-Rubin statistics presented no evidence that the chains failed to converge.

The saturated model, including the intercept, main effects, and all two-way and three-way interactions, was introduced in equations (1.1), (1.2), and (1.3). Posterior densities for the regression coefficients and random effect standard deviation are given in Figure 3.3. The data appear to have updated the diffuse normal priors on the regression coefficients. Most densities coincide, featuring similar shapes, variance, and location. ICP ZIP densities of the intercept (β_1), vaccine (β_2), and vaccine/T cell interaction (β_4) terms are shifted right. The ICP ZIP model also gives smaller estimates of the random effect standard deviation.

In Figure 3.4, we present 95% confidence intervals of the lower limit frequentist model (black) and and credible intervals of the Bayesian lower limit, ICP, and ICP ZIP models (red, green, and orange, respectively) are presented. Intervals for the lower limit models approximately coincide for all coefficients. The maximum likelihood estimate (0.42) of the random effect standard deviation is near the center of all credible intervals for this parameter. Because the models agree that a significant vaccine, protein, T cell interaction is present, we may only interpret the intercept term of this model. Here, the intercept represents the effect at baseline levels of the three binary covariates, i.e., the effect of CONS vaccine for CD4 counts and Nef protein. Although the lower limit models indicate that β_1 is not significant, ICP and ICP ZIP suggest marginal sigificance.

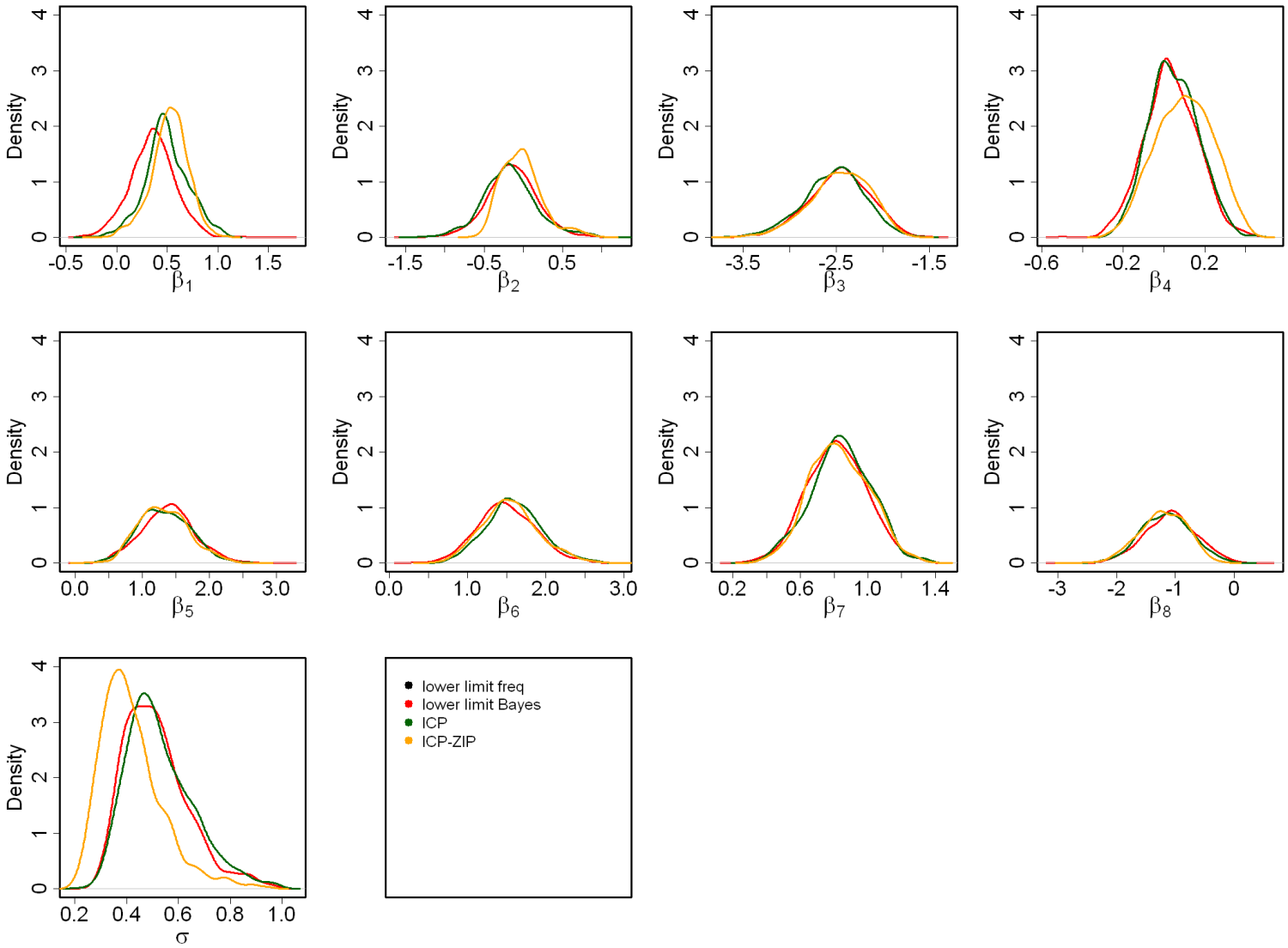

Figure 3.3: Posterior densities of regression coefficients for HIV example data

Estimated Poisson rates (Table 3.1 and Figure 3.5) are highest for ICP ZIP for most vaccine, T cell, and protein combinations. Rates estimated in the simulation study (Figures 4.3 and 4.8) are biased when all binary covariates are equal to 1. This corresponds to the case of CD8 Gag counts in the Mosaic group, so interpretation of this effect is problematic. On average, the estimated response rates of subjects treated in the Mosaic vaccine tend to be slightly higher than the CONS vaccine groups.

We separate CD4 and CD8 T cell data and analyze them under the data models introduced in equations (1.6) and (1.7), respectively. Each model includes an intercept ($\beta_{4,1},\ \beta_{8,1}$), vaccine ($\beta_{4,2},\ \beta_{8,2}$) and protein ($\beta_{4,3},\ \beta_{8,3}$) effects, and interaction terms ($\beta_{4,4},\ \beta_{8,4}$). Posterior densities (Figure 3.6) approximately coincide for all regression coefficients except $\beta_{4,1}$ where ICP and ICP ZIP densities are shifted right. Credible intervals (Figure 3.7) for all models find that the interaction between vaccine and protein effects is significant for CD4 and insignificant for CD8. All models agree that the CD8 intercept, associated with Nef counts in the CONS group, is also significant. Only the ICP ZIP model finds that the CD4 intercept is marginally significant. Poisson rates estimated from the separate CD4 and CD8 models (Figure 3.8 and Table 3.1) are similar to the rate estimates in the saturated model.

Finally, we separate the data according to T cell and protein types as in equations (1.12), (1.13), (1.14), and (1.15). Intercepts ($\beta_{4g,1},\ \beta_{4n,1},\ \beta_{8g,1},\ \beta_{8n,1}$) represent CONS vaccine effect, and a single main effect ($\beta_{4g,2},\ \beta_{4n,2},\ \beta_{8g,2},\ \beta_{8n,2}$) is the predicted increase or decrease in effect for the Mosaic group relative to CONS. Curves in the posterior density plots (Figure 3.9) feature more separation than the saturated and CD4/CD8 models. ICP ZIP posteriors tend to be narrower and have modes located at higher densities and shifted right of the other curves. For the CD4 Gag group, ICP interval estimates (Figure 3.10) give stronger evidence of a significant intercept and vaccine effects, suggesting nonzero CONS effect and non-trivial advantage for Mosaics over CONS. Although the models suggest no vaccine group differences for the CD4 Nef

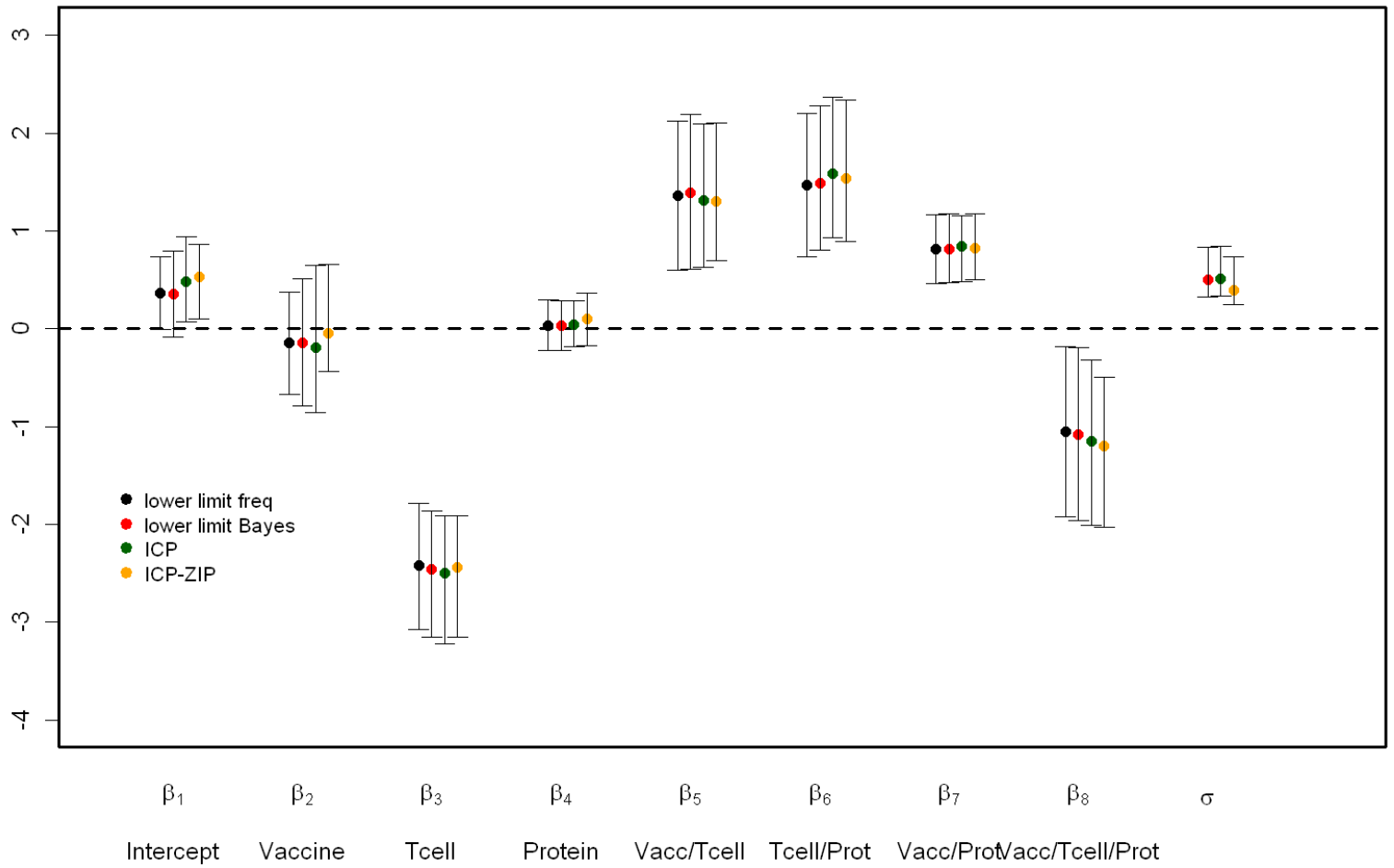

Figure 3.4: 95% credible intervals of regression coefficients for full HIV example data, calculated from posterior medians and equation (1.1)

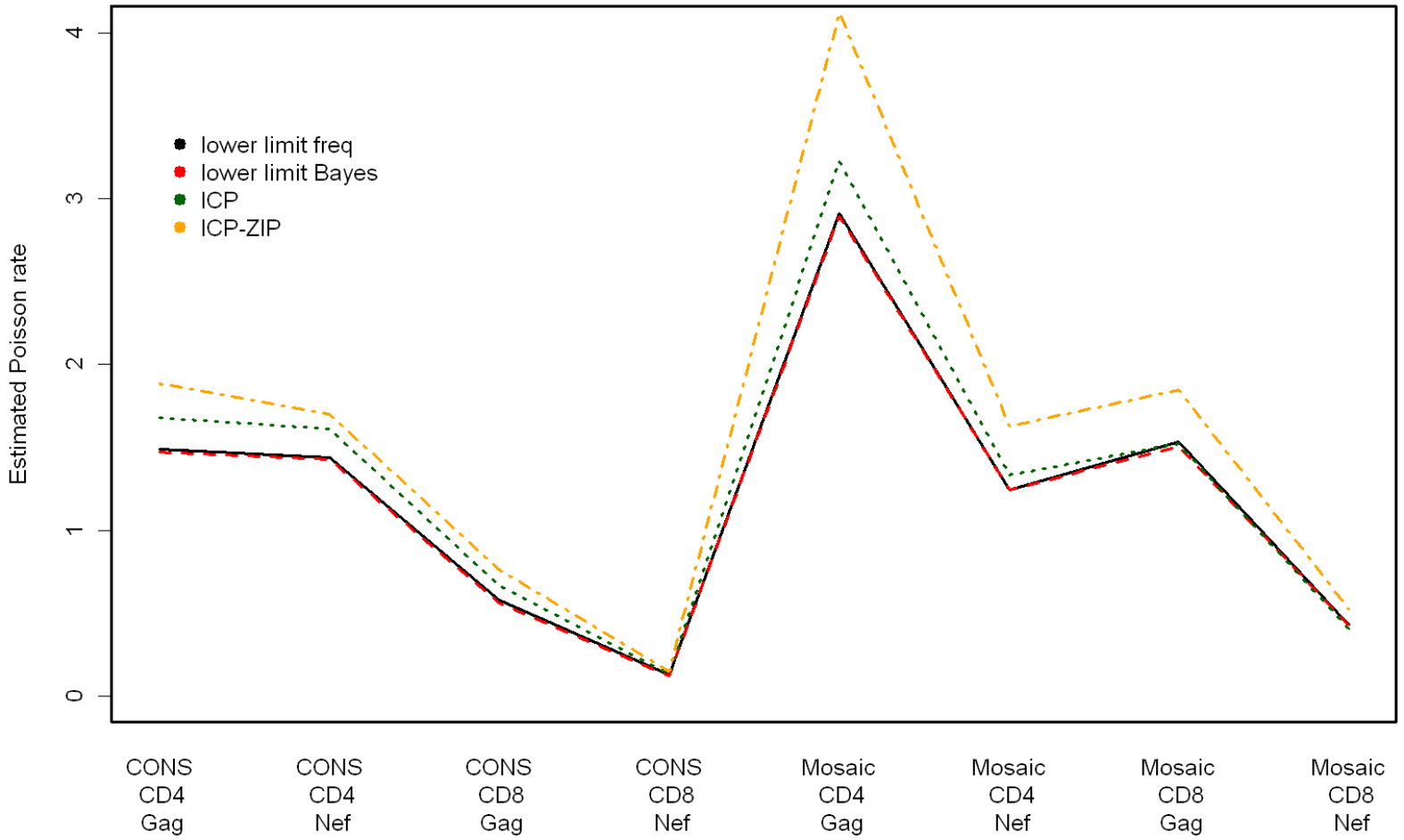

Figure 3.5: Poisson rates for full HIV example data, calculated from posterior medians and equation (1.1)

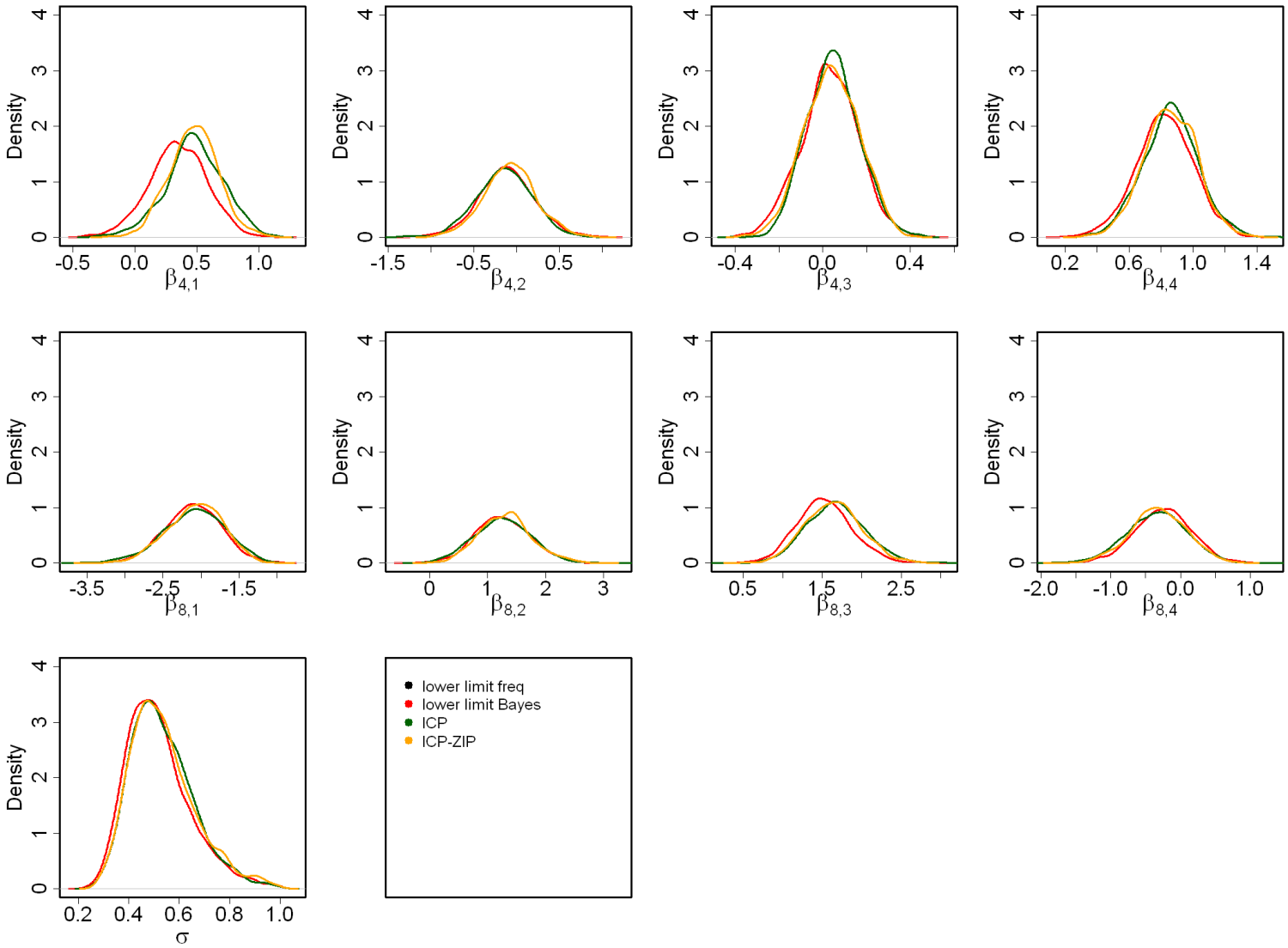

Figure 3.6: Posterior densities of regression coefficients for HIV example data

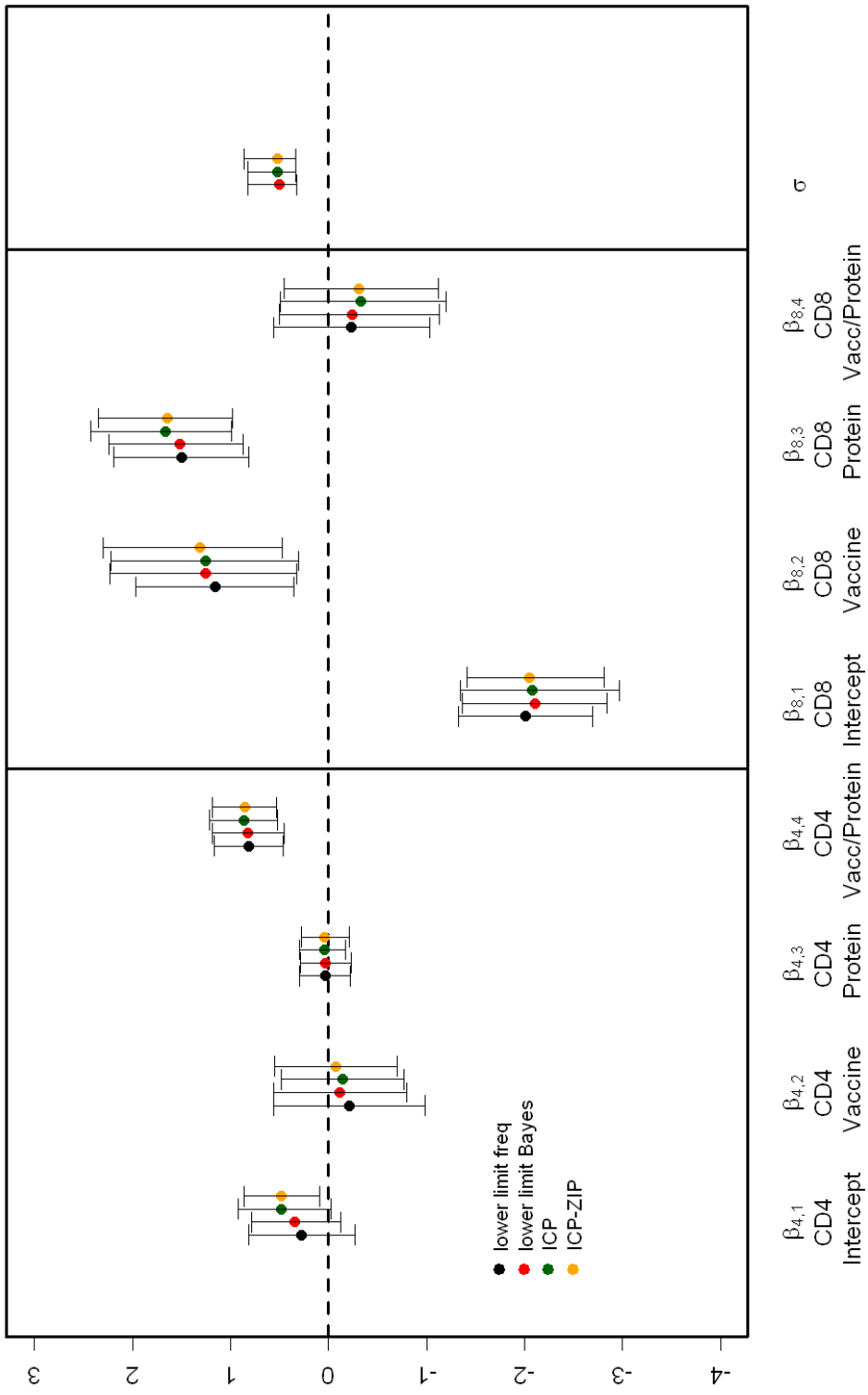

Figure 3.7: 95% credible intervals of regression coefficients for HIV example data

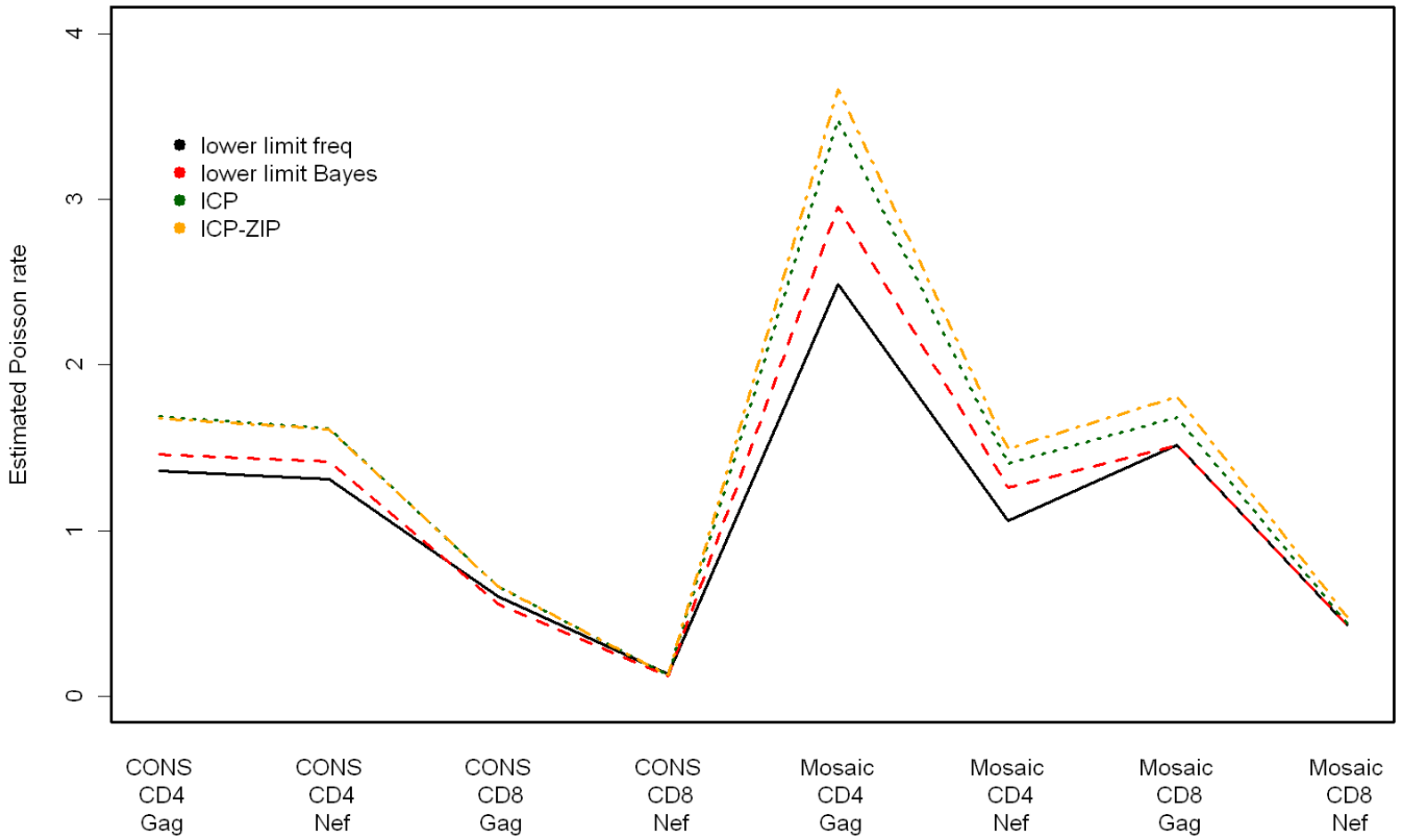

Figure 3.8: Poisson rates for separated CD4 and CD8 data, calculated from posterior medians and equations (1.6) and (1.7)

group, the ICP ZIP is the only model indicating a nonzero CONS effect. All Bayesian models find a significant advantage of Mosaics over CONS among CD8 Gag or Nef groups and nonzero CD8 Nef CONS effect. Lower limit models suggest a marginally significant effect for CD8 Gag group, while ICP and ICP ZIP models indicate that this effect is zero.

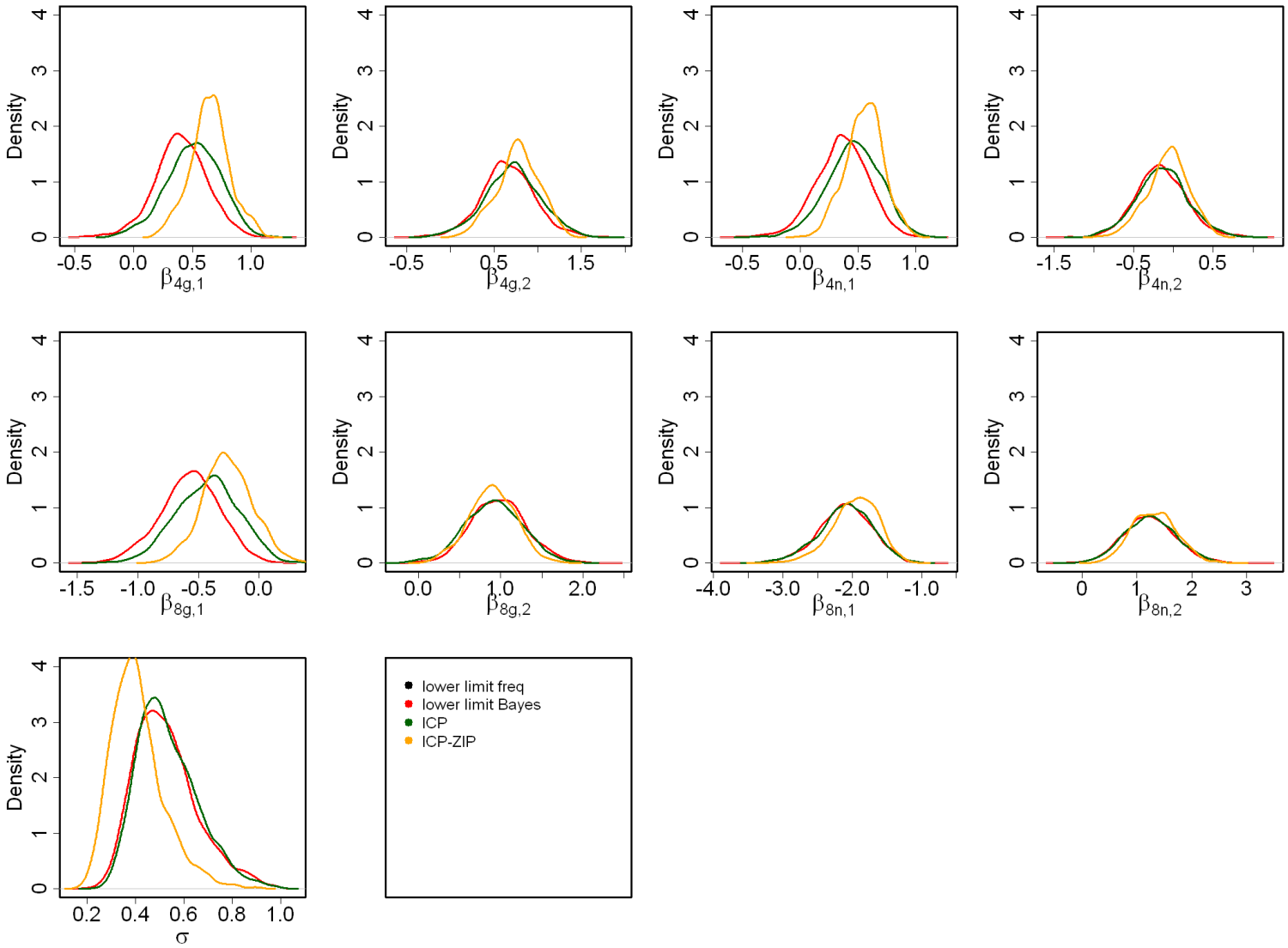

Figure 3.9: Posterior densities of regression coefficients for HIV example data

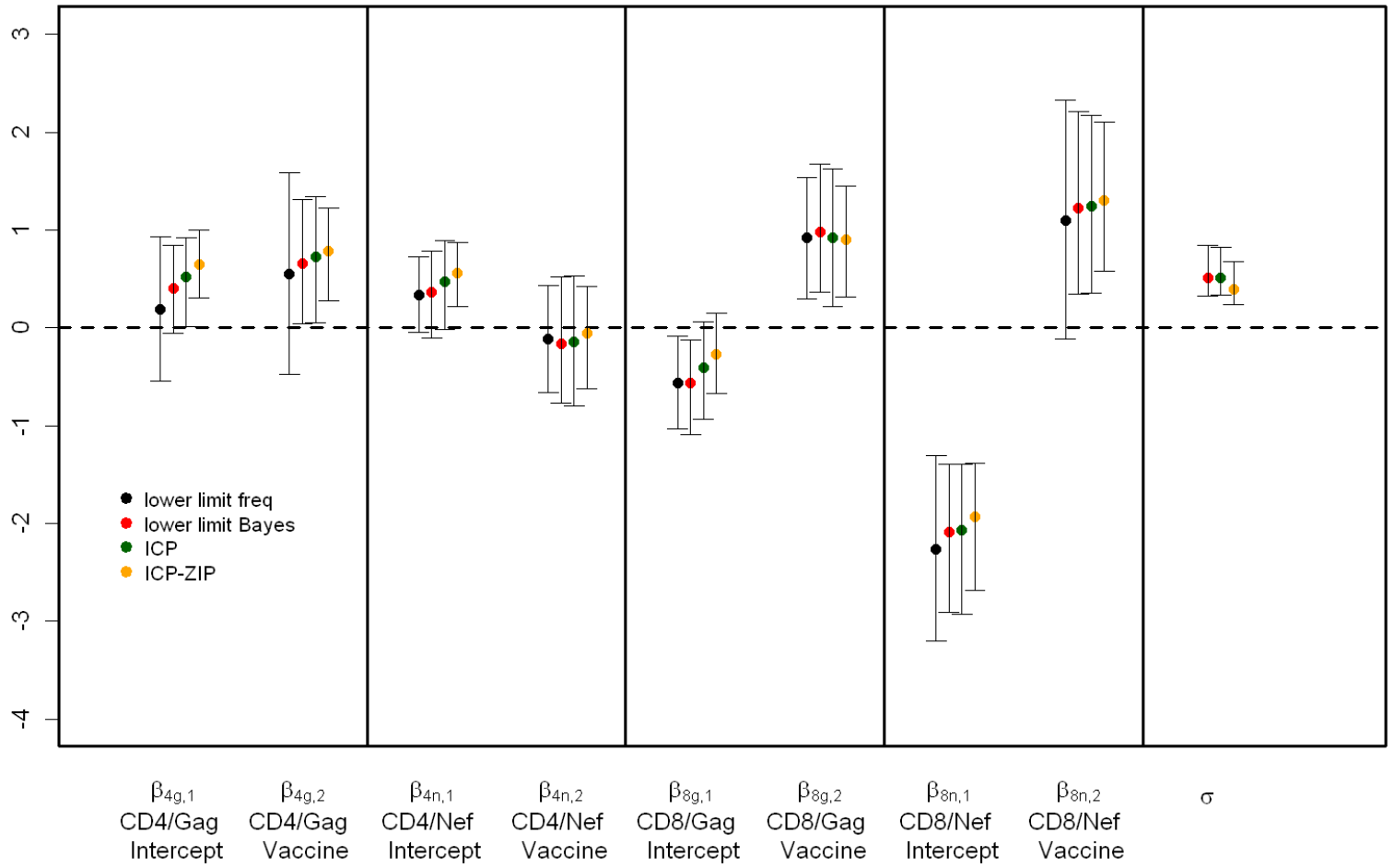

Figure 3.10: 95% credible intervals of regression coefficients for HIV example data

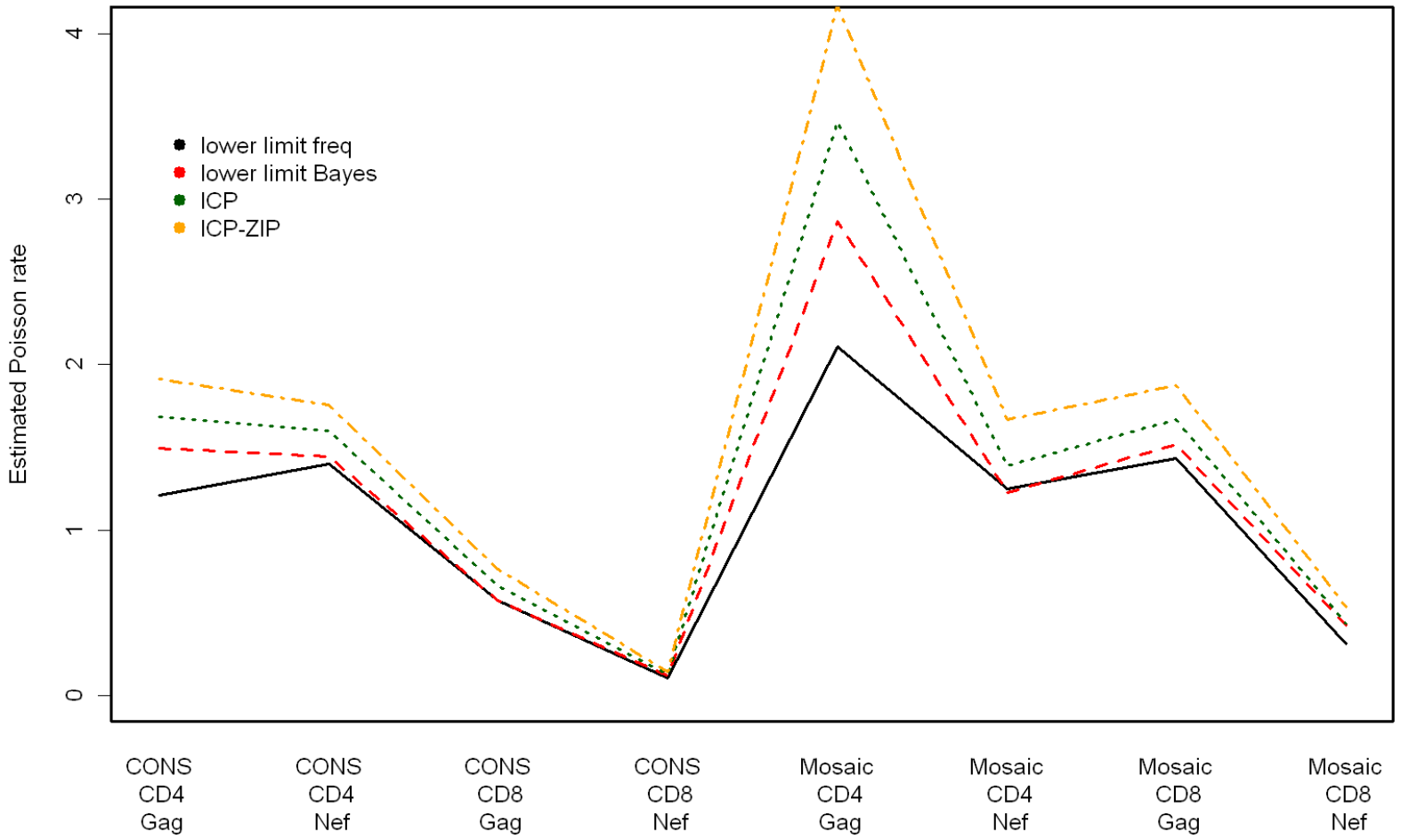

Figure 3.11: Poisson rates for separated CD4/CD8 Gag/Nef data, calculated from posterior medians and equations (1.12), (1.13), (1.14), and (1.15)

Table 3.1: Estimated Poisson rates for Santra et al. (2010) data

Model	Vaccine	Tcell	Protein	lower.lim.freq	lower.limit	ICP	ICP.ZIP
Saturated	CONS	CD4	Gag	1.49	1.47	1.68	1.89
	CONS	CD4	Nef	1.44	1.43	1.61	1.70
	CONS	CD8	Gag	0.58	0.56	0.67	0.76
	CONS	CD8	Nef	0.13	0.12	0.13	0.15
	Mosaic	CD4	Gag	2.91	2.89	3.23	4.12
	Mosaic	CD4	Nef	1.24	1.24	1.33	1.63
	Mosaic	CD8	Gag	1.54	1.50	1.52	1.84
	Mosaic	CD8	Nef	0.43	0.43	0.41	0.52
CD4/CD8	CONS	CD4	Gag	1.36	1.46	1.69	1.68
	CONS	CD4	Nef	1.31	1.41	1.62	1.61
	CONS	CD8	Gag	0.60	0.56	0.66	0.66
	CONS	CD8	Nef	0.13	0.12	0.13	0.13
	Mosaic	CD4	Gag	2.49	2.96	3.48	3.67
	Mosaic	CD4	Nef	1.06	1.26	1.41	1.50
	Mosaic	CD8	Gag	1.52	1.52	1.69	1.81
	Mosaic	CD8	Nef	0.43	0.43	0.44	0.48
Gag/Nef	CONS	CD4	Gag	1.21	1.49	1.69	1.91
	CONS	CD4	Nef	1.40	1.44	1.60	1.76
	CONS	CD8	Gag	0.57	0.57	0.66	0.76
	CONS	CD8	Nef	0.10	0.12	0.13	0.14
	Mosaic	CD4	Gag	2.11	2.87	3.47	4.17
	Mosaic	CD4	Nef	1.25	1.23	1.39	1.67
	Mosaic	CD8	Gag	1.43	1.52	1.67	1.87
	Mosaic	CD8	Nef	0.31	0.42	0.44	0.53

3.6.1 HIV Example Summary

In our analysis of the interval censored HIV example data, conclusions from the ICP ZIP model more closely resemble the results in the Barouch et al. (2010) study of Mosaic and CONS vaccines and the theoretical comparison of these vaccines in Fischer et al. (2007) than the results from the lower limit analysis in Santra et al. (2010). As in those studies, we find that most CONS effects are nonzero, and the Mosaic vaccine group generally gives a significantly higher response than CONS. In using this model, however, we were forced to discard data associated with any subjects missing CD4/CD8 distinction. Selection bias is still a concern. We have also not addressed the fact that we disregarded the total (CD4 + CD8) T cell counts which are known for all subjects. Given the current parameterization of the model, these totals are sums of dependent Poissons variates. Future analyses of this dataset will incorporate this dependence into the likelihood. The Gibbs sampler in `OpenBUGS` does not allow distribution parameters to depend on missing data, so we are not able to incorporate missing data into the ICP or ICP ZIP model. We plan to use the full conditionals of the parameters to implement a Gibbs sampler for ICP and ICP ZIP models and apply them to the HIV example data.

CHAPTER FOUR

Interval Censored Poisson Regression: Simulation Study

The purpose of this simulation study is to compare three candidate Poisson regression models of imprecise measurements with a model of the precise counts. The Bayesian model of the lower limits of the intervals was introduced in Santra et al. (2010) and presented in Section 2.2. In Chapter 3, we proposed two Poisson regression models of interval censored count data, one including an adjustment for zero inflation in the lower limit (ICP ZIP), and one without (ICP). In the study, we demonstrate that the lower limit model gives imprecise and inaccurate parameter estimates for the design points tested. We also discuss whether adjusting for zero inflation improves performance of the interval models. Datasets generated in the study include either low, moderate, or extreme censoring widths, allowing us to investigate the impact that observed interval widths have on model performance.

This chapter is organized as follows. In Section 4.1, we describe a simulation study which formally compares models of the lower limits of interval-censored observations to the interval censored Poisson regression models developed in Chapter 3. Next, in Section 4.2, we describe criteria under which we evaluate the models of interest. Finally, in Section 4.3, we present results from the fixed and mixed effect models.

4.1 Simulation Design and Implementation

We begin this section with a description of the method used to simulate interval-censored Poisson data for our simulation study. We also introduce the criteria used to evaluate and compare the proposed interval censoring models with the model of the lower limits. The structure of the data in the simulation study is, by design, similar to

the HIV example data presented in the introduction. Let Y_i be a precise count arising from a Poisson(θ_i) process with binary covariates $x_{1i}, x_{2i} \in \{0, 1\}$, where $i = 1, ..., n$ index the n subjects in the study. Then the log-linear fixed effects model of the data is

$$\log \theta_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{1i} x_{2i}. \tag{4.1}$$

We evaluate the models in the context of small sample size ($n = 30$), as small samples are common in HIV vaccine studies.

In a mixed effect model, precise counts $Y_{ij} \sim$ Poisson(θ_i) are associated with subject i and replication j, where $j = 1, ..., R$, and R is the number of replications per subject. Introducing the random effect $\nu_i \sim \mathrm{N}(0, \sigma_\nu^2)$ to (4.1), as in Section 1.1, yields the following log-linear mixed effect model.

$$\log \theta_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{1i} x_{2i} + \nu_i. \tag{4.2}$$

We set $R = 10$ in the simulated data sets.

The set of covariate vectors and the true value of the regression coefficient vector completely specify the collection of expectations, θ_i, for all subjects. These θ_i are used to simulate data according to the underlying Poisson process. Data are simulated using regression vectors that feature (a) only insignificant (zero) effects or (b) all significant positive or negative effects of various sizes. Table 4.6 gives the collection of regression coefficient vectors considered in our study. In this chapter, we present results for simulated data with $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$. Results for other coefficient vectors were similar, with a few exceptions noted in Section 4.3.3. Summary plots for the remaining coefficients are presented in the appendix. Mixed effect model data were generated with random effect standard deviation $\sigma_\nu = 0.5$.

The mean of the simulated process must be considered in the selection of possible interval widths for censoring. For example, an interval of width 9 (e.g., $j = 1$, $k = 10$) for a process with mean 5 would render the censored datum useless. By contrast, observing an interval of the same width, wtih $j = 81$ and $k = 90$ for a process with

mean 85 may still be informative. Therefore, in our simulations, we varied the censoring interval width according to the size of the true value of θ. Specifically, let w denote the maximum amount of interval censoring allowed, computed as

$$w = \lceil m \cdot min_i \{\theta_i\} \rceil = \lceil m \cdot min_i \{\mathbf{x}_i'\boldsymbol{\beta}\} \rceil, \tag{4.3}$$

where m is a scalar multiple of the mean parameter in $\{0.5, 1.5, 2.5\}$ and $\lceil \cdot \rceil$ denotes the ceiling function. These three values define low, moderate, and extreme censoring widths, respectively.

Our interval censored data mimic those that arise from an underlying Poisson process. We simulate intervals given maximum allowable interval width w and mean parameter θ as follows. First, generate a random variate $x \sim \text{Poisson}(\theta)$. Next, generate a pair (h_1, h_2) of positive integers such that $0 \leq h_1 + h_2 \leq w$, $h_1 \in \{0, ..., min(w, x)\}$, and $h_2 \in \{0, ..., (w - h_1)\}$. Then the simulated interval censoring limits are $j = x - h_1$ and $k = x + h_2$.

Recall that the interval-censored Poisson likelihoods introduced in Pruszynski (2010) exclude left-censoring ($j = 0$) due to irregularity. Poisson counts which are exactly observed at $x = 0$ are not possible in such a construction and are excluded. Sampling according to an algorithm which excludes left censored intervals generates values from a zero-truncated Poisson process. We graphically illustrated (Figure 3.2) regularity of the likelihood for θ given a set of censored intervals including at least one observation which is not left-censored. Since approximately 61% of the observed intervals in the HIV example data are left-censored (or exactly equal to zero), our focus is on the models' relative performance in the presence of left-censoring. Thus, our simulated datasets may include left censored intervals and/or values exactly observed to be zero, but these cases may comprise no more than 70% of any dataset.

To ensure that the distributions of parameter estimates are directly comparable, we model the same datasets using the four models. Posterior means serve as point estimates for the parameters, and we also record 95% credible intervals and posterior standard deviations for each parameter. The models are analyzed using MCMC methods as implemented in the `OpenBUGS` software. The code is provided in Appendix A.4. We discard 1000 thinned burn-in iterations and save 5000 thinned samples after burn-in with a thinning rate of 25. The prior distributions for the regression coefficients were independent normal distributions centered at 0 with reasonably diffuse variances (100).

Model selection is complicated by the different representations of the data in the models we consider. In particular, we compare four models with different dependent variables (i.e., lower limits, precise count, and two models of the observed intervals). Most model selection criteria, including Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Deviance Information Criteria (BIC), and Bayes factors, require that the compared models are used to analyze the same set of numeric values. Although our two proposed ICP models are comparable using any of these criteria, this limitation motivates our need to explore other ways to compare and contrast various features of the four models simultaneously.

Point estimates are compared using smoothed kernel densities of the $N = 500$ recorded posterior medians under each model for each regression coefficient estimated. In these graphs, approximately symmetric, unimodal curves with small variance suggest more precise measurement. Distribution centers which are closer to the true value used to simulate the data are consistent with higher accuracy.

Bias is another feature considered in our assessment of model accuracy. For any given model and regression coefficient β_k, the bias is $\beta_k - \hat{\beta}_k$, where $\hat{\beta}_k$ is the posterior median associated with β_k under the model. We illustrate the distribution of biases

using boxplots. Boxplots that are centered near zero and with low interquartile difference are associated with more accurate measurement. Another measure of accuracy is the expected bias which we estimate using the simulation average bias. Average bias is summarized by error bars which are centered at the simulation mean bias and extend to 1 average standard deviation of the bias in either direction.

Coverage of the 95% credible intervals is graphically summarized with shaded error bars. These bars are centered at the simulation means of the posterior medians and extend to the simulation means of the lower and upper 95% credible intervals. Shaded boxes around the average posterior means and interval limits extend to 1 standard deviation of these averages in either direction.

We also compare the significance of the terms in the log-linear model. Significance is a frequentist concept which is adapted to test the hypothesis of a nonzero regression coefficient in Bayesian models. In this manuscript, we use the term *significant* to refer to effects that are more likely to be nonzero given the model and data. To assess significance of a given term, we record the proportion of times that zero is included in 95% credible intervals corresponding to that term.

4.3 Simulation Study Results

In this section, we discuss the relative performance of lower limit, precise count, and two interval censoring models. We analyze data generated according to the log-linear fixed or mixed effect models in equations (4.1) and (4.2), respectively.

4.3.1 Fixed Effect Model Results

In Figures 4.1 - 4.5, we graphically summarize analyses of data generated from the fixed effect model in equation (4.1). The first row of plots are the smoothed kernel densities of $N = 500$ posterior medians estimating regression coefficients β_1, β_2, β_3,

and β_4 respectively for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$. The curves correspond to the lower limit (red), interval-censored (green), zero-inflated interval-censored (orange), and precise (blue) models. Plots on this row are associated with data that feature the lowest amount of interval censoring in our study. The maximum interval width computed according to equation (4.3) with $m = 0.5$ is $w = 1$, so intervals in the simulated datasets are allowed to have widths of 0 (i.e., precise measurement) or 1. Vertical solid lines mark the true value of the coefficient used to simulate the data.

Because the model features only binary covariates, the values of the coefficients are interpreted relative to baseline levels of the two pairs of covariate groups. For example, the intercept, β_1, measures the baseline effect when covariate values associated with β_2 and β_3 are zero. Of the four models, the mode of the density of simulation posterior medians in the lower limit model of β_1 is furthest away from the true value of 0.5. The closer proximity of the ICP mode to the truth indicates that this parameter tends to be more accurately estimated under that model. The distribution of estimates from the ICP ZIP model approximately coincide with the estimates from the model of precise counts. This suggests that the ICP-ZIP estimates for the baseline treatment group closely resemble those we would have obtained had we precisely measured the counts. The distributions of estimates in the interval model are tighter around their modes, indicating more precise estimation.

Posterior median distributions of main effect coefficients β_2 and β_3 and interaction coefficient β_4 are given in the second, third, and fourth plots in the first row, respectively. The curves do not feature the separation that we observe in distributions of β_1 esttimates. Modes of these distributions are located at the true value of 0.5. Each set of curves appears to exhibit a similar degree of variability, although the lower limit distribution appears to have slightly heavier tails. For this case of low interval censoring, these plots suggest that imprecise measurement does not have a significant impact on estimation of the main effects and interactions.

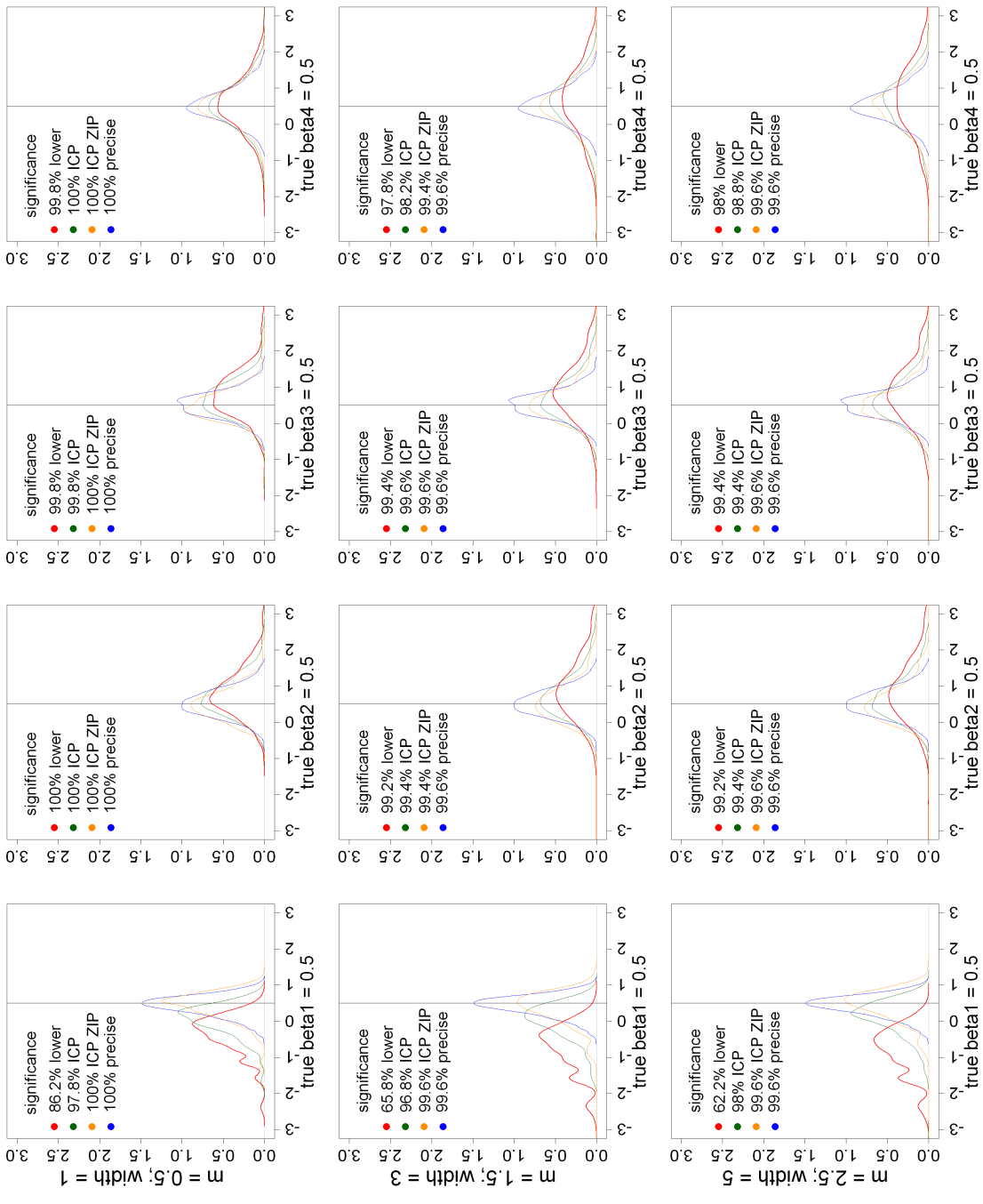

Figure 4.1: Posterior medians for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in fixed effects model

Tables in the upper left corners of each plot give the percentage of replications in the simulation in which a given coefficient is found to be significant in any of the four models. Recall that significance of an effect is determined by the the exclusion of 0 from 95% credible intervals. Intercepts were significant in 100% of the simulated datasets under the precise and interval models, while only 86.2% were significant in the lower limit model. The distributions of β_2, β_3, and β_4 estimates are similar, so it is not surprising that the four models tend to agree on significance (99.8% - 100%) for these terms.

Rows 2 and 3 of Figure 4.1 display the posterior median distributions under the four models given moderate ($m = 1.5$) and extreme ($m = 2.5$) amounts of censoring in the simulated data, with maximum interval widths of $w = 3$ or $w = 5$, respectively. Because the precise counts include no censoring (and, hence, no consideration of increased interval width), the blue curves are identical for $m = 0.5, 1.5$, and 2.5. Increasing interval widths in the simulated data appears to increase the variability in the estimates in the interval and lower limit models. The effect is most substantial for the lower limit model; the estimates are considerably less precise and less accurate, and the distribution appears to be skewed left. The interval censoring model estimates appear to be relatively stable when the intervals are wider.

The likelihood of significance is also affected by widening the interval widths in the simulated data. The significance proportions for β_1 decrease from 86.2% to 66% and 62.4%, while 98.4% - 100% of the main effects and interaction term were found to be significant under the precise and interval models. Again, we see that the ICP-ZIP estimates are more consistent with the precise model estimates than those from the lower limit and ICP models.

Bias of parameter estimates for data generated with $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ is graphically summarized by the plots in Figure 4.2 for low (left column), moderate (middle), and extreme censoring (right). Above each parameter on the horizontal axis,

we plot a group of error bars of the average bias (top row) or boxplots of raw bias scores (bottom row) observed in the four models. The error bars are centered at the arithmetic means of the $N = 500$ bias scores and extend in either direction to one sample standard deviation of the average biases. As in the posterior median distribution curves, graphs associated with the precise count model are identical for all amounts of censoring. Means and medians of bias for all coefficients in the precise models are approximately zero. For the smallest interval widths, bias for β_1 is mostly negative and highly variable. Intercept bias is smallest in the ICP and ICP-ZIP models. The model controlling for zero inflation appears to have a slight advantage in accuracy over the ICP model and has a distribution that more closely resembles precise count model bias. Bias distrbutions for β_2, β_3, and β_4 are similar in the four models, with centers near zero and exhibiting high variability. Increasing interval widths has more of an effect on bias for the intercept term in the lower limit model than all others terms and models; bias becomes more negative as widths become larger.

Using equation (4.1), we calculate four possible Poisson rates given four binary covariate pairs and true $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ (Table 4.1). We also use posterior medians $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)'$ from each model to compute the rates estimated in each run. Boxplots of the $N = 500$ results (Figure 4.3) are grouped by covariate value $(x_1 = 0, x_2 = 0)$, $(x_1 = 0, x_2 = 1)$, $(x_1 = 1, x_2 = 0)$, $(x_1 = 1, x_2 = 1)$ for low, moderate, and extreme censoring. Bias in the estimated Poisson rates exhibit similar behavior as observed in the bias of the estimated intercept terms. The horizonal line marking the true Poisson rate passes through the center of the rates estimated by the precise count model. Most rates estimated from the lower limit model are negatively biased, and increasing interval widths appears to introduce more bias in the rate estimates. Estimated rates in the interval censoring models are closer to the precise count model. The ICP ZIP rates most closely resemble the precise model rates for all pairs except for $(x_1 = 1, x_2 = 1)$ where they approximately coincide. Increasing interval widths appears

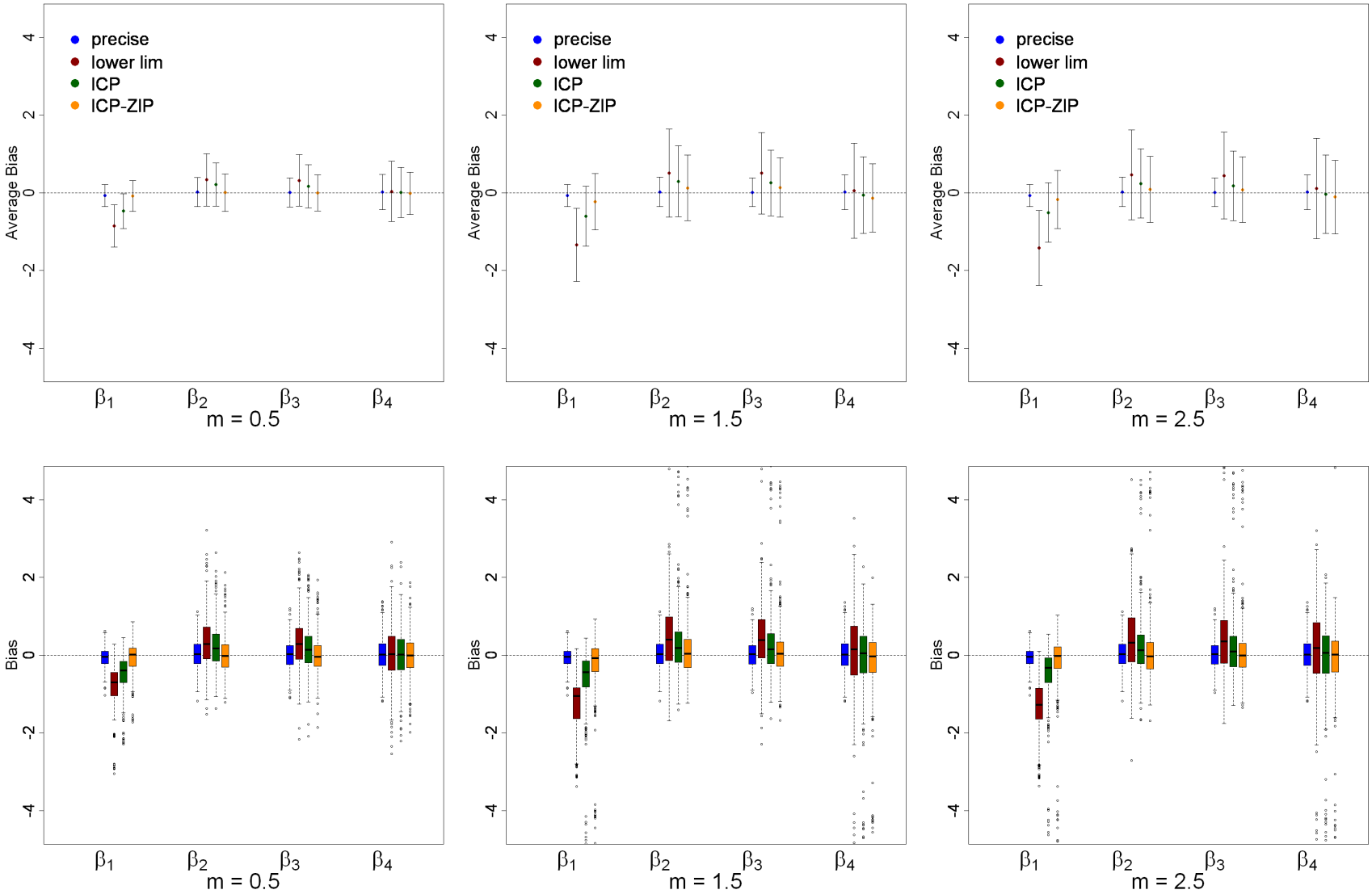

Figure 4.2: Bias for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

Table 4.1: True poisson rates for $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

X_1	X_2	$\theta \mid X_1, X_2, \boldsymbol{\beta}$
0	0	1.65
0	1	2.72
1	0	2.72
1	1	7.39

to have no significant effect in the ICP and ICP ZIP estimated rates: estimates are similar for low, moderate, and extreme levels of censoring.

Bias boxplots and posterior median densities are used to compare the overall behavior of the distributions of these quantities. In contrast, a scatterplot of posterior medians of the precise model against those computed from the lower limit and interval models (Figure 4.4) directly compares the estimates of the models from the same simulated dataset. For any given model, an exact linear relationship (blue line) in these plots indicates exact agreement with the corresponding estimate from the precise count model. Overall, posterior medians in the lower limit and interval models appear to linearly increase with those in the precise model, yet estimates from all models vary considerably around the true value of 0.5. Three fairly distinct clusters of points appear in the intercept plots (left) for all interval censoring widths. As the interval widths increase, correspondence between precise and lower limit posterior medians decreases. Although most ICP and ICP ZIP intercept posterior medians are close to the identity line, the disparity between the estimates and the precise model estimates increases with wider interval censoring widths. For non-intercept terms, the plots do not reveal any difference in predictive correspondence to the precise count model in the lower and interval models.

Each group of shaded error bars in Figure 4.5 summarizes coverage of the $N = 500$ 95% posterior credible intervals for the regression coefficient indicated on the horizontal axis and the four models in our study. These bars are centered at the

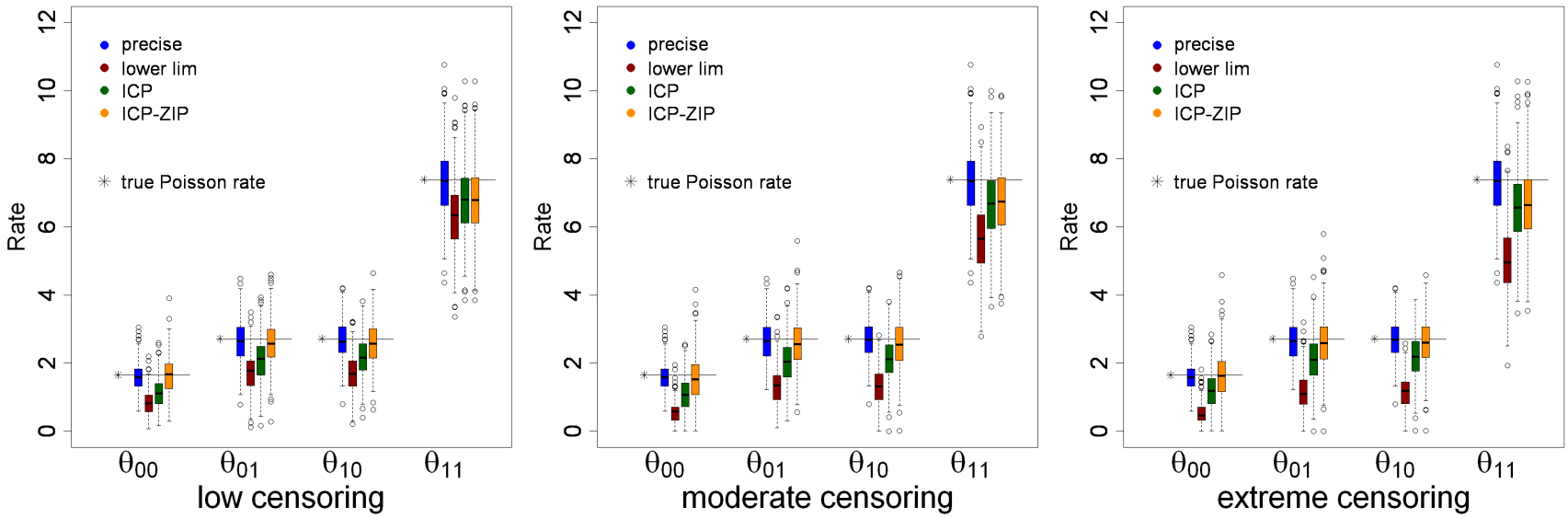

Figure 4.3: Poisson rates for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

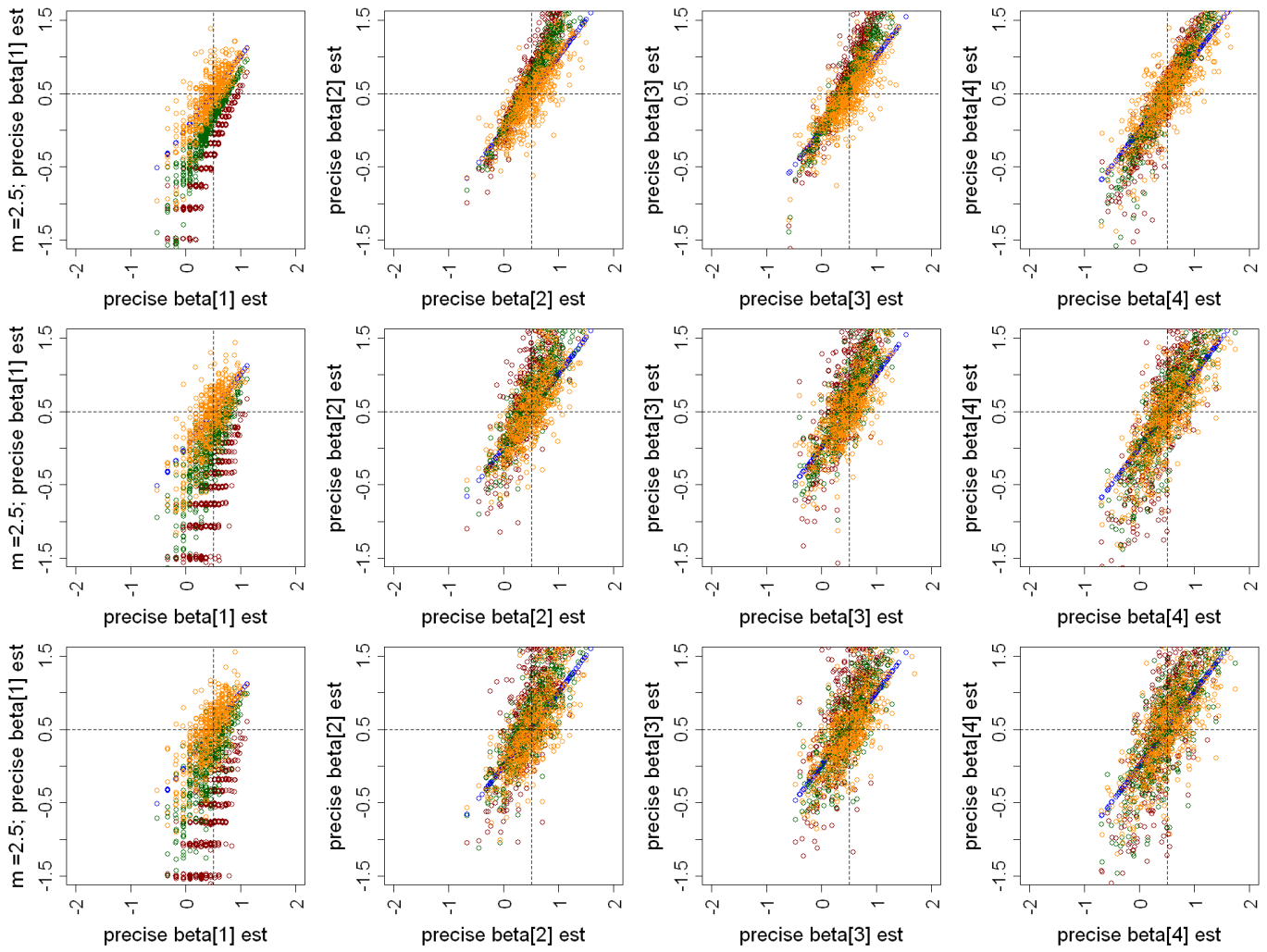

Figure 4.4: Lower limit and interval model estimates plotted against precise count model estimates for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

simulation means of the posterior medians and extend to the simulation means of the lower and upper 95% credible intervals. Shaded boxes around the average posterior medians and interval limits extend to 1 standard deviation of these averages in either direction. Most error boxes for the credible interval limits overlap with the error boxes of the posterior medians, indicating a considerable amount of variability in the interval estimates. On average, the true coefficient value is within 1 standard deviation of the average posterior median for non-intercept terms. Precise and ICP ZIP posterior medians are closer to the true value. The true intercept is higher than most of the upper 95% credible limits in the lower limit model, indicating poor coverage. ICP and ICP ZIP models appear to have greater coverage, but this may be attributed to the wider credible interval widths.

We also summarize 95% credible interval coverage and credible interval widths in Tables 4.2 and A.2. Credible intervals from the precise model have the lowest average widths and provide approximately 93-95% coverage for all terms. Average credible interval widths increase as count intervals in the simulated datasets incease in width for the lower and ICP models, especially for the intercept term. Although the lower limit credible intervals are wide, coverage is still very poor, decreasing from 39% with low censoring to 15 or 13% for wider censoring widths. Coverage is very high (97-99%) in the ICP ZIP models for all censoring widths, but this may be attributed to wider credible interval widths.

4.3.2 Mixed Effects Model Results

Analyses of data generated from the mixed effects model (4.2) are summarized in Figures 4.6 - 4.7 via plots of the posterior medians and biases observed under the four models. For $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$, posterior median and bias distributions for all coefficients (Figures 4.6 and 4.7) feature less dispersion than the analogous fixed effects models (Figures 4.1 and 4.2). All distributions of precise count point estimates

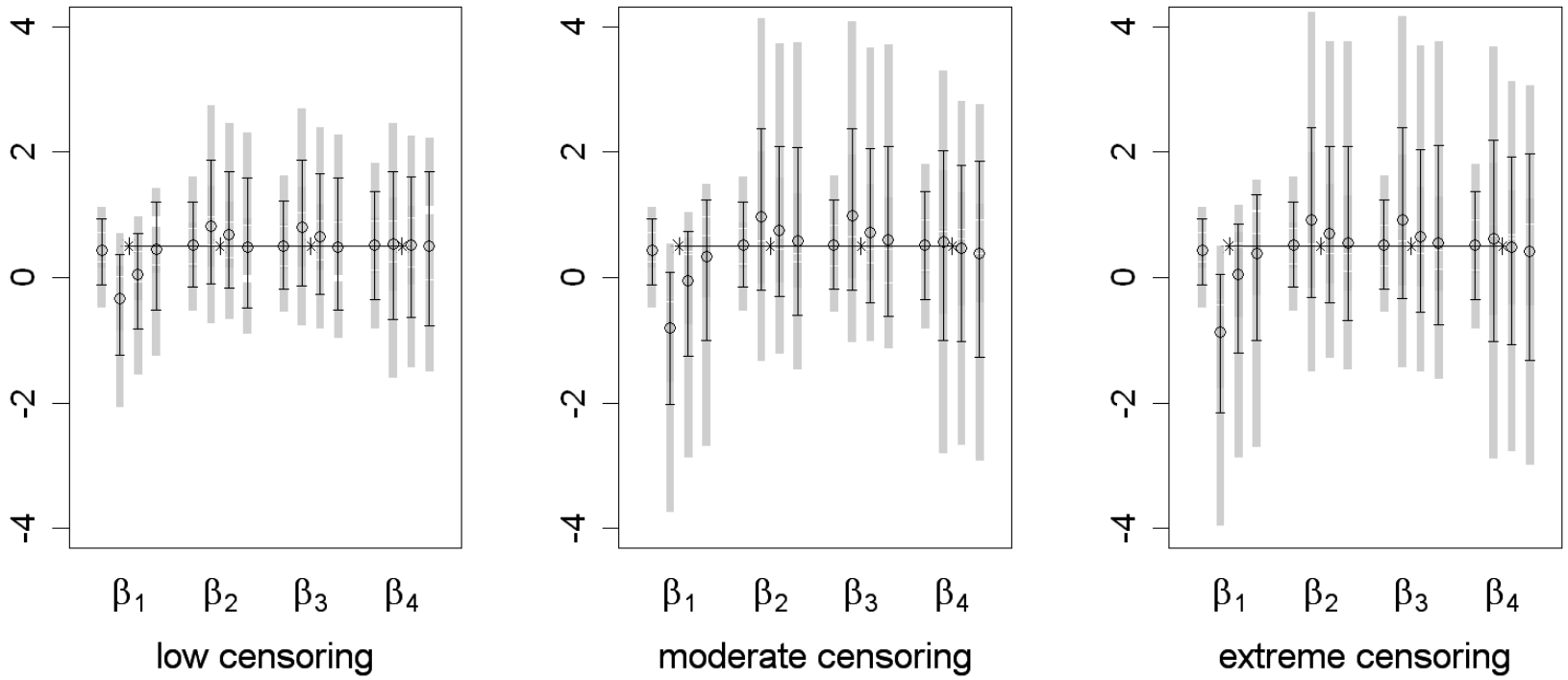

Figure 4.5: Summary of the coverage of 95% credible intervals for regression coefficient estimates for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

are centered at their true value and exhibit least variability. Estimates of β_1 under the lower limit model are negatively biased and are furthest away from the model of true counts for all interval widths. Although the ICP model is consistently less biased than the lower limit model, the ICP intercept biases are centered near zero. Controlling for zero inflation appears to be effective in yielding estimates which are more consistent with the precise count model. Estimates of β_2, β_3, and β_4 are fairly accurate under each of the four models. Surprisingly, the ICP estimates tend to be the least precise. Increasing interval widths does not appear to have much of an affect on precision or accuracy in any of the models. Nearly 100% of the estimates for each parameter under the precise and interval models are significant. Wider interval widths were associated with lower significance proportions in the lower limit model.

The estimated Poisson rates (Figure 4.8) for $(x_1 = 0, x_2 = 0)$, $(x_1 = 0, x_2 = 1)$, and $(x_1 = 1, x_2 = 0)$ behave similarly for mixed and fixed effects. Precise and ICP model rates are centered at their true value, while the ICP and lower limit model rates are negatively biased. However, for $(x_1 = 1, x_2 = 1)$, even for the precise model, estimated rates appear to be biased. This result is surprising due to the low bias in the estimation of the coefficients in all models. We suspect that this discrepancy may be attributed to induced priors. Specifying the multivariate normal for $\boldsymbol{\beta}$ in (2.2) induces a univariate normal prior on the log rate including only fixed effects, as $\log \theta_i | \sigma_B^2 \sim N_q(\mathbf{x}_i' \boldsymbol{\beta}, \sigma_\nu^2)$. Addition of the normally distributed random effect similarly induces a univariate normal prior with larger variance, as in $\log \theta_i | \sigma_B^2, \sigma_\nu^2 \sim N_q(\mathbf{x}_i' \boldsymbol{\beta}, \sigma_B^2 + \sigma_\nu^2)$, and the uniform prior on the random effect standard deviation introduces even more uncertainty.

Pairwise scatterplots of the posterior medians of the precise mixed model against those in the lower limit and interval models for mixed effects are given in Figure 4.9. Lower limit point estimates appear to linearly increase with precise estimates for all terms and all simulated interval widths. The lower limit scatterplot of intercept es-

timates is parallel to the identity line, suggesting that intercept estimates under this model are biased by approximately the same amount. We observe this relationship across all censoring widths. The linear relationship between the precise count model and ICP or ICP ZIP model estimates is weak; their scatterplots have elliptictal shapes approximately centered near the intersection of the reference lines at 0.5. This suggests that although ICP and ICP ZIP model estimates are approximately centered at the true values, parameters in these models are not precisely estimated.

In the plots summarizing average coverage of the $N = 500$ simulated 95% credible intervals (Figure 4.10), most β_2, β_3, and β_4 intervals appear to cover their true values on average. The true value is within 1 standard deviation of the average posterior medians of the precise and interval models for these terms, while only the β_4 interval in the lower limit model is near the truth. ICP and lower limit estimates are negatively biased, with the true value located inside or above the upper limit error box, respectively, with more negative bias in the lower limit model as censoring interval widths increase. As in the fixed effects model, the ICP and ICP ZIP error boxes overlap.

Tables 4.4 and 4.5 also summarize the coverage of the mixed effects models. Credible intervals in the precise count model are narrowest and cover the true value aproximately 100% of the time. Lower limit credible intervals are widest, and coverage is good (94-100%) for non-intercept terms and poor (0-4%) for the intercept. ICP coverage for non-intercept terms range from 75% to 93%, but intercept coverage drops to 58-75%. ICP ZIP coverage is higher for non-intercepts (89-95%) and intercepts (94-95%). These tables do not suggest any systematic decrease in coverage as censoring interval widths increase.

4.3.3 Discussion

In the simulation study, we analyze simulated interval censored Poisson count datasets with fixed or mixed regression models of either the precise counts, lower limits, or the intervals. The results of this chapter correspond to data simulated from true $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ and estimate regression coefficients with medians of their respective posterior distributions. For main effects β_2 and β_3 and interaction term β_4, average bias is approximately zero with high variance for all fixed effects models and very low variance for mixed effects. ICP and ICP ZIP estimates of the intercept term β_1 feature less bias and more precision than the lower limit estimates. Using a mixture ICP model to control for zero inflation appears to effectively pull estimates away from 0 and closer to the true value. Estimates from the interval models appear to be more precise than lower limit models. Coverage is highest in the ICP ZIP models, but this may be attributed to wider interval widths. Average estimated Poisson rates in the ICP and ICP ZIP fixed effects models are closer to the true value than the lower limit model, but these estimates tend to be negatively biased. We find similar results in the mixed model for $(x_1 = 0, x_2 = 0)$, $(x_1 = 0, x_2 = 1)$, and $(x_1 = 1, x_2 = 0)$. However, for $(x_1 = 1, x_2 = 1)$, estimated rates appear to be biased, even given the precise counts.

We arrive at the same conclusions regarding bias, precision, and coverage for most of the other design points in our simulation study (Table 4.6). One notable exception is the fixed effects model for data generated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$. Results for fixed and mixed effects models for this $\boldsymbol{\beta}$ are given in Appendix Section A.5. Posterior median densities are very flat, suggesting that there was little updating from the diffuse multivariate normal prior on $\boldsymbol{\beta}$ even when counts are precisely measured. Poisson rates do not appear to be affected by the extreme variation in the posterior medians. Here, 95% credible interval coverage is low despite the extremely wide credible interval widths. We do not find similar issues in the analogous mixed effects models. This may indicate that the log-linear fixed effects model is not a good fit to data having a

negative baseline effect. To avoid the problems associated with a negative intercept, we may consider coding the binary variables so that the hypothesized baseline is positive. We will continue to investigate the properties of the intercept term and develop a modification to the ICP models which address the issues that arise when the baseline effect is negative.

Table 4.2: Coverage for lower limit and precise count fixed effects models and true $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$

Model	Term	m	Truth	Avg.width	SD.width	Coverage
Lower limit	β_1	0.50	0.50	1.60	0.49	0.39
	β_2	0.50	0.50	1.97	0.46	0.81
	β_3	0.50	0.50	2.01	0.45	0.87
	β_4	0.50	0.50	2.36	0.43	0.87
	β_1	1.50	0.50	2.12	1.29	0.15
	β_2	1.50	0.50	2.58	1.37	0.78
	β_3	1.50	0.50	2.57	1.27	0.81
	β_4	1.50	0.50	3.02	1.34	0.84
	β_1	2.50	0.50	2.20	1.39	0.13
	β_2	2.50	0.50	2.71	1.49	0.81
	β_3	2.50	0.50	2.72	1.47	0.84
	β_4	2.50	0.50	3.21	1.55	0.86
Precise count	β_1	0.50	0.50	1.05	0.16	0.93
	β_2	0.50	0.50	1.35	0.14	0.94
	β_3	0.50	0.50	1.41	0.15	0.95
	β_4	0.50	0.50	1.72	0.14	0.95
	β_1	1.50	0.50	1.05	0.16	0.93
	β_2	1.50	0.50	1.35	0.14	0.93
	β_3	1.50	0.50	1.41	0.15	0.95
	β_4	1.50	0.50	1.72	0.13	0.95
	β_1	2.50	0.50	1.05	0.16	0.93
	β_2	2.50	0.50	1.35	0.14	0.93
	β_3	2.50	0.50	1.41	0.15	0.95
	β_4	2.50	0.50	1.72	0.13	0.95

Table 4.3: Coverage for ICP fixed effects models and true $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$

Model	Term	m	Truth	Avg.width	SD.width	Coverage
ICP	β_1	0.50	-0.50	6.13	4.03	0.53
	β_2	0.50	0.50	7.53	4.48	0.68
	β_3	0.50	0.50	7.91	4.66	0.69
	β_4	0.50	0.50	9.13	4.88	0.73
	β_1	1.50	-0.50	6.13	4.03	0.53
	β_2	1.50	0.50	7.53	4.48	0.68
	β_3	1.50	0.50	7.91	4.66	0.69
	β_4	1.50	0.50	9.13	4.88	0.73
	β_1	2.50	-0.50	6.50	4.03	0.46
	β_2	2.50	0.50	7.96	4.35	0.61
	β_3	2.50	0.50	8.61	4.73	0.60
	β_4	2.50	0.50	9.85	4.80	0.62
ICP ZIP	β_1	0.50	0.50	1.72	0.57	0.99
	β_2	0.50	0.50	2.06	0.52	0.99
	β_3	0.50	0.50	2.11	0.56	0.98
	β_4	0.50	0.50	2.46	0.51	0.99
	β_1	1.50	0.50	2.25	1.50	0.97
	β_2	1.50	0.50	2.68	1.55	0.97
	β_3	1.50	0.50	2.71	1.44	0.98
	β_4	1.50	0.50	3.12	1.50	0.99
	β_1	2.50	0.50	2.32	1.54	0.98
	β_2	2.50	0.50	2.77	1.56	0.98
	β_3	2.50	0.50	2.85	1.60	0.98
	β_4	2.50	0.50	3.28	1.61	0.98

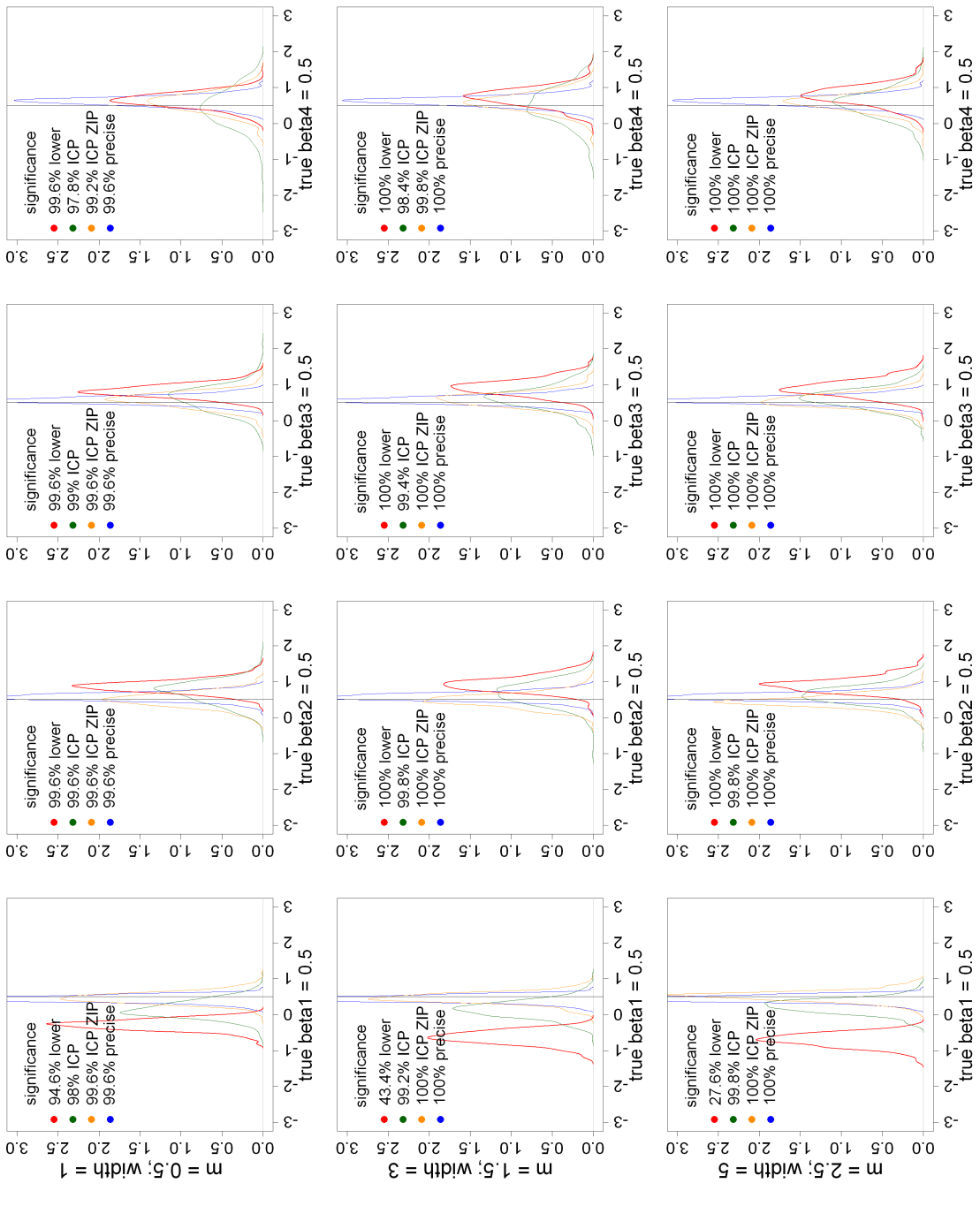

Figure 4.6: Posterior medians for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in mixed effects model

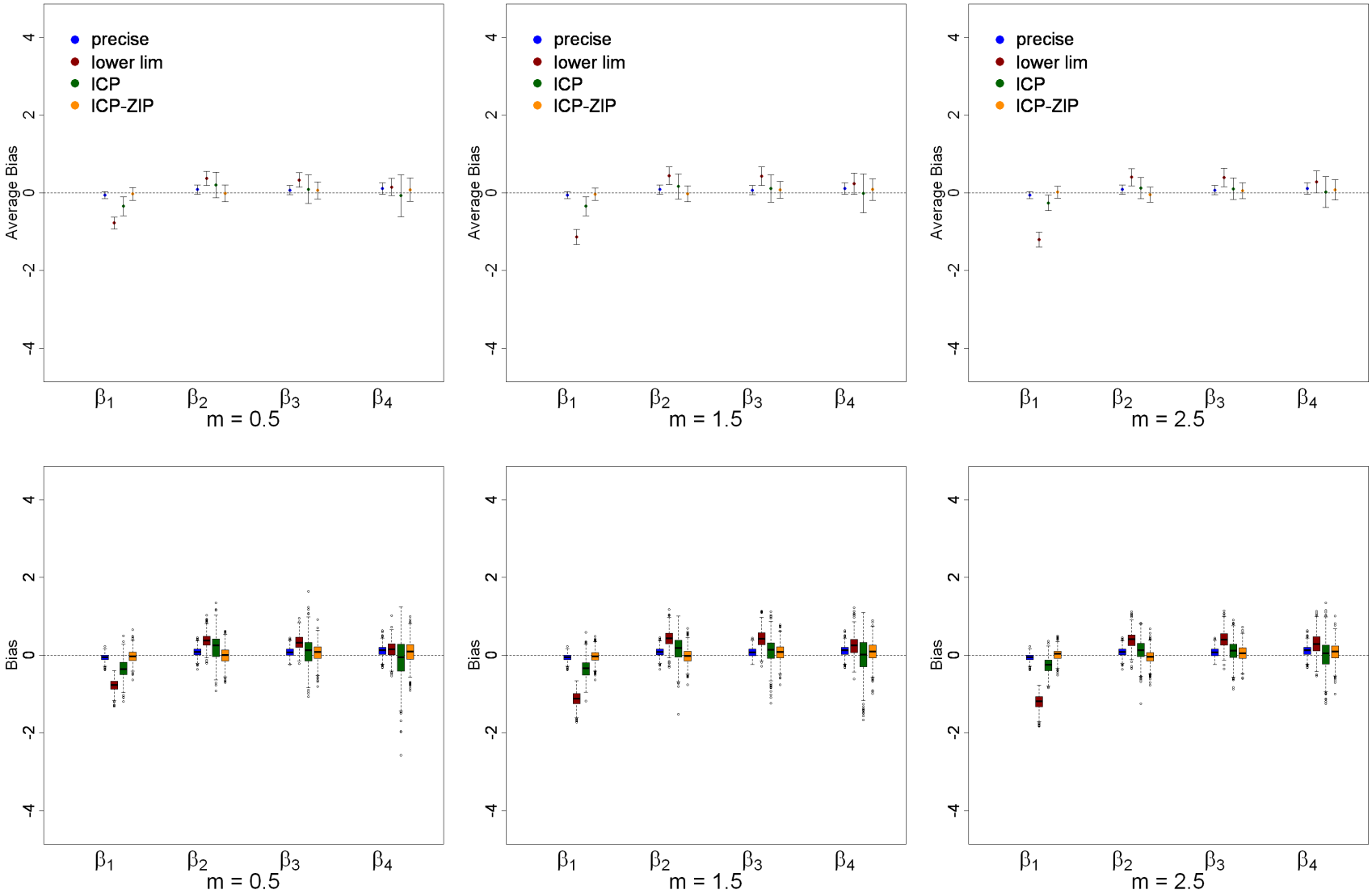

Figure 4.7: Bias for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in mixed effects model

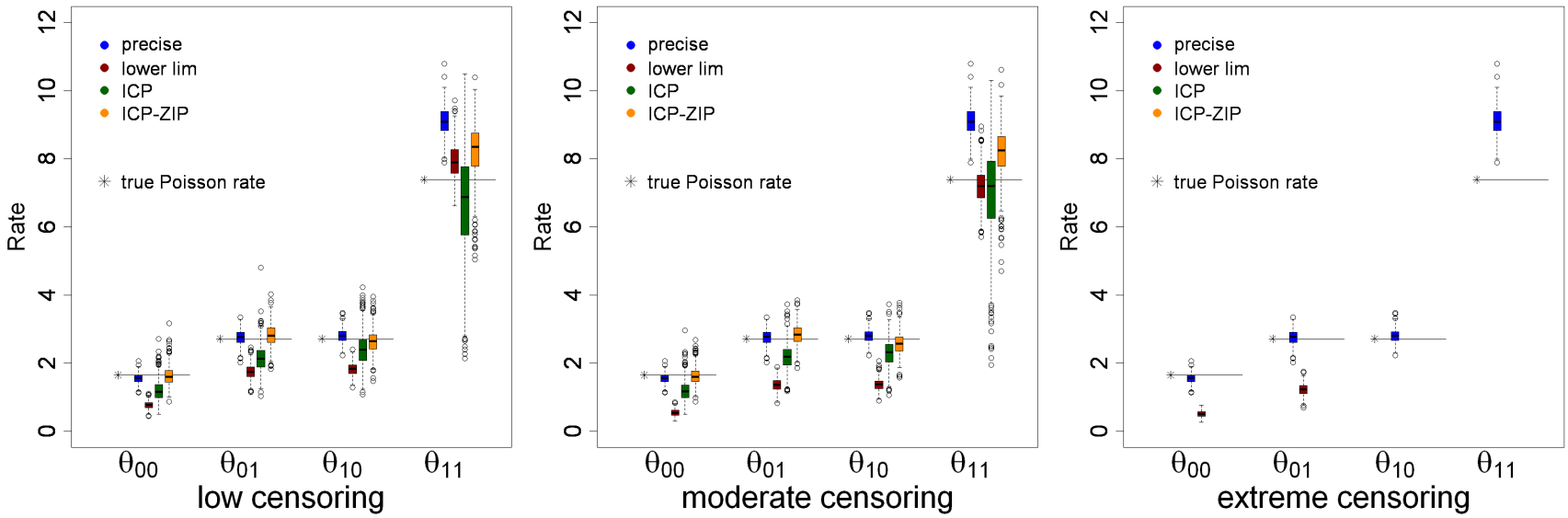

Figure 4.8: Poisson rates for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in mixed effects model

Figure 4.9: Lower limit and interval model estimates plotted against precise count model estimates for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in mixed effects model

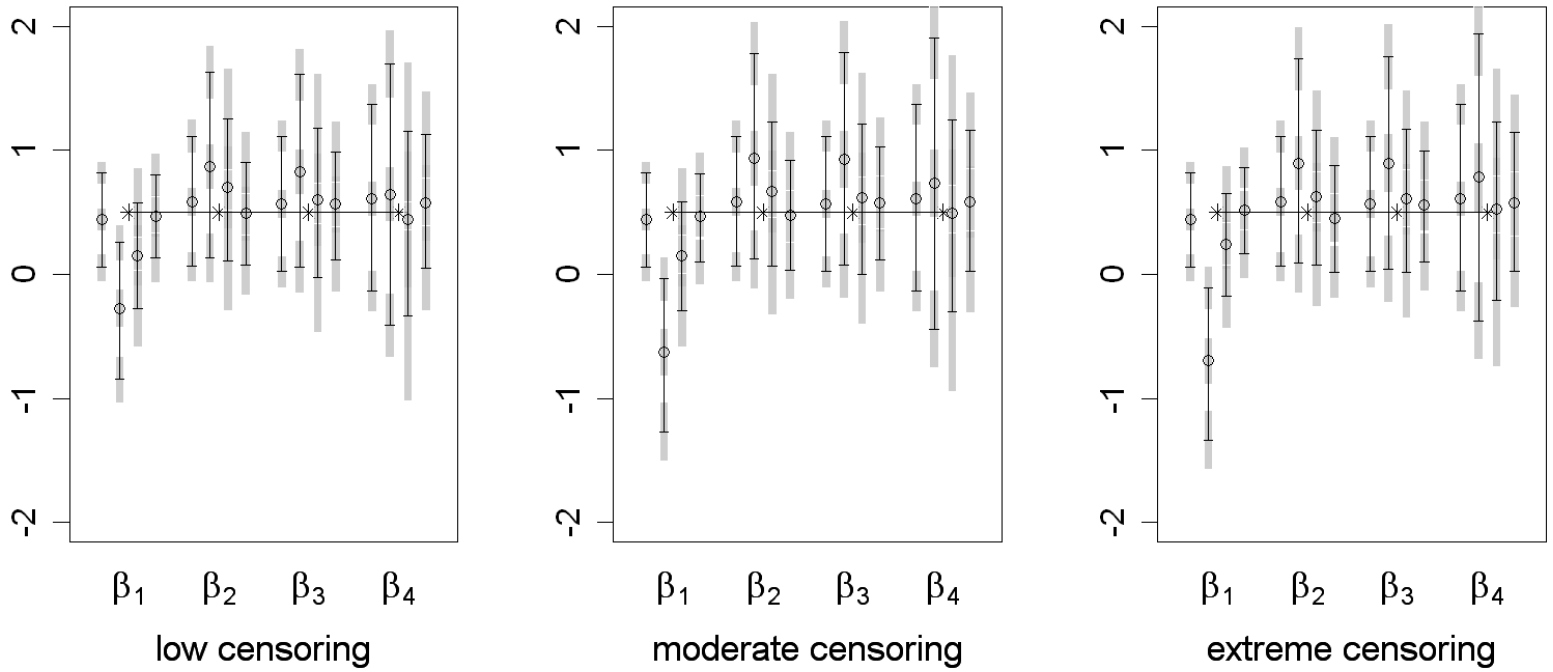

Figure 4.10: Summary of the coverage of 95% credible intervals for regression coefficient estimates for data simulated from $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$ in mixed effects model

Table 4.4: Coverage for lower limit and precise count mixed effect models and true $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$

Model	Term	m	Truth	Avg.width	SD.width	Coverage
Lower limit	β_1	0.50	0.50	1.10	0.12	0.04
	β_2	0.50	0.50	1.49	0.17	0.96
	β_3	0.50	0.50	1.55	0.18	0.97
	β_4	0.50	0.50	2.11	0.26	1.00
	β_1	1.50	0.50	1.23	0.14	0.00
	β_2	1.50	0.50	1.66	0.19	0.94
	β_3	1.50	0.50	1.72	0.21	0.95
	β_4	1.50	0.50	2.34	0.31	1.00
	β_1	2.50	0.50	1.22	0.14	0.00
	β_2	2.50	0.50	1.65	0.20	0.96
	β_3	2.50	0.50	1.71	0.21	0.95
	β_4	2.50	0.50	2.31	0.30	1.00
Precise count	β_1	0.50	0.50	0.76	0.07	1.00
	β_2	0.50	0.50	1.05	0.10	1.00
	β_3	0.50	0.50	1.09	0.10	1.00
	β_4	0.50	0.50	1.50	0.14	1.00
	β_1	1.50	0.50	0.76	0.07	1.00
	β_2	1.50	0.50	1.05	0.10	1.00
	β_3	1.50	0.50	1.09	0.10	1.00
	β_4	1.50	0.50	1.50	0.14	1.00
	β_1	2.50	0.50	0.76	0.07	1.00
	β_2	2.50	0.50	1.05	0.10	1.00
	β_3	2.50	0.50	1.09	0.10	1.00
	β_4	2.50	0.50	1.50	0.14	1.00

Table 4.5: Coverage for ICP mixed effects models and true $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5)'$

Model	Term	m	Truth	Avg.width	SD.width	Coverage
ICP	β_1	0.50	0.50	0.86	0.31	0.58
	β_2	0.50	0.50	1.14	0.42	0.82
	β_3	0.50	0.50	1.20	0.47	0.86
	β_4	0.50	0.50	1.48	0.65	0.79
	β_1	1.50	0.50	0.88	0.28	0.60
	β_2	1.50	0.50	1.16	0.43	0.87
	β_3	1.50	0.50	1.21	0.43	0.87
	β_4	1.50	0.50	1.54	0.66	0.87
	β_1	2.50	0.50	0.82	0.26	0.75
	β_2	2.50	0.50	1.08	0.36	0.90
	β_3	2.50	0.50	1.15	0.41	0.93
	β_4	2.50	0.50	1.43	0.57	0.90
ICP ZIP	β_1	0.50	0.50	0.67	0.18	0.94
	β_2	0.50	0.50	0.83	0.23	0.92
	β_3	0.50	0.50	0.87	0.25	0.92
	β_4	0.50	0.50	1.08	0.36	0.89
	β_1	1.50	0.50	0.71	0.18	0.94
	β_2	1.50	0.50	0.89	0.24	0.95
	β_3	1.50	0.50	0.92	0.26	0.94
	β_4	1.50	0.50	1.14	0.34	0.92
	β_1	2.50	0.50	0.70	0.18	0.95
	β_2	2.50	0.50	0.86	0.23	0.94
	β_3	2.50	0.50	0.90	0.24	0.95
	β_4	2.50	0.50	1.12	0.34	0.94

Table 4.6: Vectors $\boldsymbol{\beta}$ used in the simulation study

$(0, 0, 0, 0)'$
$(1, 1, 1, 1)'$
$(-1, 1, 1, 1)'$; $(1, -1, 1, 1)'$; $(1, 1, -1, 1)'$; $(1, 1, 1, -1)'$
$(0.5, 1, 1, 1)'$; $(1, 0.5, 1, 1)'$; $(1, 1, 0.5, 1)'$; $(1, 1, 1, 0.5)'$
$(0.5, 0.5, 0.5, 0.5)'$
$(-0.5, 0.5, 0.5, 0.5)'$; $(0.5, -0.5, 0.5, 0.5)'$; $(0.5, 0.5, -0.5, 0.5)'$; $(0.5, 0.5, 0.5, -0.5)'$
$(0.25, 0.5, 0.5, 0.5)'$; $(0.5, 0.25, 0.5, 0.5)'$; $(0.5, 0.5, 0.25, 0.5)'$; $(0.5, 0.5, 0.5, 0.25)'$

We use the log link function because it is the canonical link function for Poisson generalized linear mixed models, and the log-linear Poisson regression has been extensively studied. In future work, we will investigate whether other link functions may be more appropriate for the interval censored Poisson models. We will also determine how sensitive these results are to sample sizes different from $n = 30$. Also, in contrast to the single maximum interval width assigned per dataset in the current study, later studies will consider varying maximum interval widths across subjects.

APPENDICES

APPENDIX A

Appendices

A.1 Vaccine Designs

Human immunodeficiency virus (HIV) is a retrovirus that attacks the immune system. When it enters the body, it begins to disable the body's immune system by using the body's aggressive immune responses to the virus to infect, replicate and kill immune system cells. Gradual deterioration of immune function is what leads to acquired immunodeficiency syndrome (AIDS).

The problem of HIV vaccine development is complicated due to extreme genotypic variability of the virus. The two main types of HIV are HIV-1 and HIV-2. HIV-1 is divided into three main groups: M (the dominating group), N (extremely rare), and O (restricted to central/west Africa). There are 9 subtypes, called clades, within Group M (A, B, C, D, F, G, H, J, K). The clades are further divided into subtypes (A1 - A3, etc). In the HIV replication process, various genetic mutations can occur. In the process of copying the virus, mistakes often occur, resulting in insertions or deletions in the HIV sequences. Two different HIV strains can also combine to form a new strain in a process known as recombination. All of these contribute to the extreme genetic diversity of HIV. As a consequence, amino acids can diverge by 15% or more in the same clade and 30% or more in different clades.

Vaccines stimulate the body's immune system to provide protection against infection or disease. Many types of HIV vaccines are being developed and are under study in various clinical trials. Simek et al. (2009) report results of a trial involving more than 16,000 adults in Thailand and a vaccine which combined two other vaccines that had previously failed. The report indicated that their proposed vaccine regimen

is safe and reduces by 31% the chance of infection with HIV. However, this vaccine was only tested against strains of HIV that are common in Thailand, so it is unclear whether this vaccine will be effective for other strains. A successful HIV vaccine design must address this global diversity.

Two vaccine strategies addressing genetic diversity have been investigated: HIV-1 global consensus envelope sequence (CON-S) and Polyvalent vaccine antigens (Mosaic). The consensus and mosaic strategies both create immunogens (i.e., substances that prompt the generation of antibodies and can cause an immune response) that resemble natural proteins resembling what the body would process in natural infections. The body cannot tell the difference between these synthetic proteins and those found in nature. We now consider the two vaccine strategies in detail and discuss how each addresses HIV genotypic diversity.

Construction of the M-group consensus (CON-S) vaccine begins with an alignment of all available HIV-1 gene sequences from group M. Usually only 2-3 genes of interest are chosen and aligned. The sequence assembled using the most prevalent amino acid at each position is the immunogen used in the CON-S vaccine. As illustrated in Figure A.1, the M-group consensus sequence is half the genetic distance from any two cross-clade sequences than they are from each other.

Fischer et al. (2007) showed that the mosaic vaccine design recognizes a higher number of epitopes than the consensus vaccine. This approach assembles proteins using a computational method that creates a vaccine which is optimized to cover the largest possible number of T-cell epitopes for a given population of HIV strains. Construction of the mosaic vaccine also begins with an alignment of a few select regions of all available HIV-1 gene sequences from group M. We artificially form a population of recombinant strains by randomly selecting pairs of strains from the original population and taking bits from the two to form a new strain, mimicking the natural recombination process. In this manner, k of these populations of recombinants are formed. Figure A.2

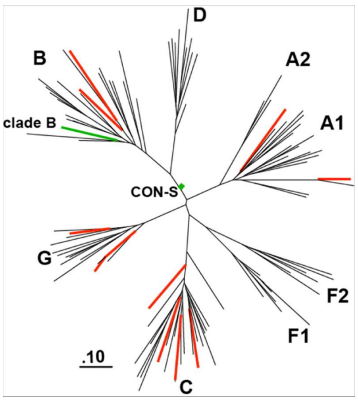

Figure A.1: A phylogenetic tree derived using a maximum likelihood approach (Olsen et al., 1994) is used to display the evolutionary behavior of HIV-1 strains diverging from a common ancestral strain, or "root". Genetic distances (or branch lengths in the tree) represent the amount of evolution (i.e., the percentage of difference between sequences) between the two nodes they connect. Longer branch lengths correspond to more highly evolved strains. In the figure, the M-group consensus sequence is the root of the tree and is half the genetic distance between sequences in different clades. Used with permission from Santra et al. (2008).

illustrates a mosaic design with $k = 4$ populations. A mosaic cocktail is assembled from the random selection of one sequence from each of the k populations. This cocktail is scored to determine the number of epitopes in the original population of strains which are covered (or recognized). An iterative process then randomly replaces sequences in the cocktail with other sequences until some maximal amount of coverage of epitopes is attained.

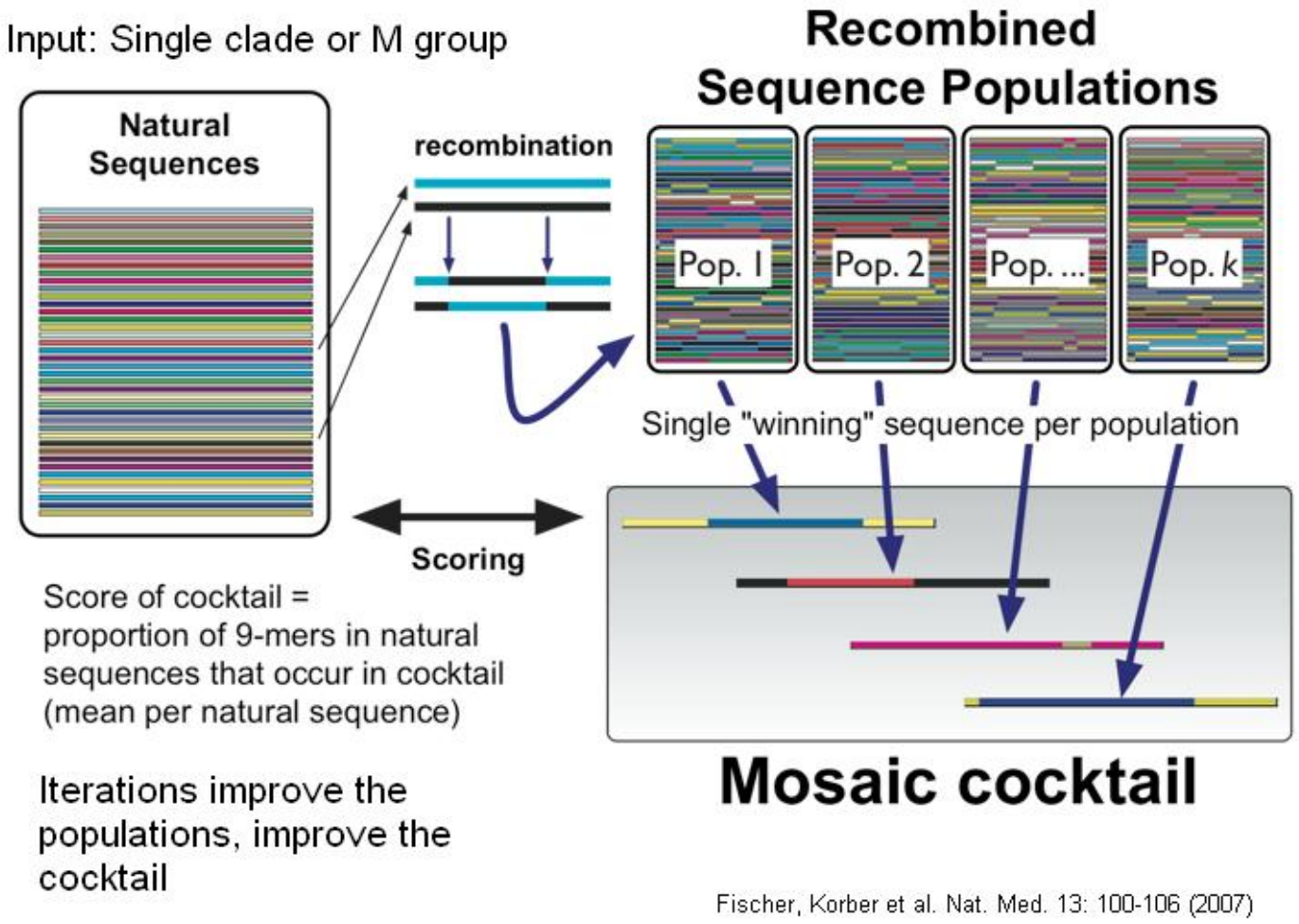

Figure A.2: Summary of Mosaic vaccine design. Used with permission from Fischer et al. (2007).

A.2 Convergence

In equation (2.2), we specified a uniform prior on random effect standard deviation σ_ν for the mixed effects model introduced in equation (1.3). An extreme amount of autocorrelation for $\tau = 1/\sigma_\nu^2$ was observed when using a less informative prior ($B_\nu = 5$), requiring a thinning rate of 40 to achieve convergence. It was proposed that decreasing B_ν from 5 to 1 may improve convergence. Also, because the posterior density of β_3 was centered at approximately 0, we questioned whether the convergence rate could also be improved by making the prior on β_3 more informative (i.e., centered at 0 with higher precision).

The plots of the posterior densities, traces, and autocorrelations of 5 models are shown in the following graphs. First, we considered the original model as previously specified with a thinning rate of 40 (Figures A.3 - A.5) or 25 (Figures A.6 - A.8). The next three models featured a thinning rate of 25 and

- prior $\sigma_\nu \sim U(0.01, B_\nu = 1)$ (Figures A.9 - A.11),
- prior $\beta_3 \sim N(0, \tau_3 = 0.144^2)$, where τ_3 is the prior precision of β_3 (Figures A.12 - A.14), or
- priors $\sigma_\nu \sim U(0.01, B_\nu = 1)$ and prior $\beta_3 \sim N(0, \tau_3 = 0.144^2)$ (Figures A.15 - A.17).

The model featuring higher precision for β_3 (Figure A.14) does not appear to decrease the amount of autocorrelation from what we observed in the original model at thinning rate 25 (Figure A.8). However, the models including a smaller upper bound on the prior for σ_ν (Figures A.11 and A.17) both seem to significantly improve convergence.

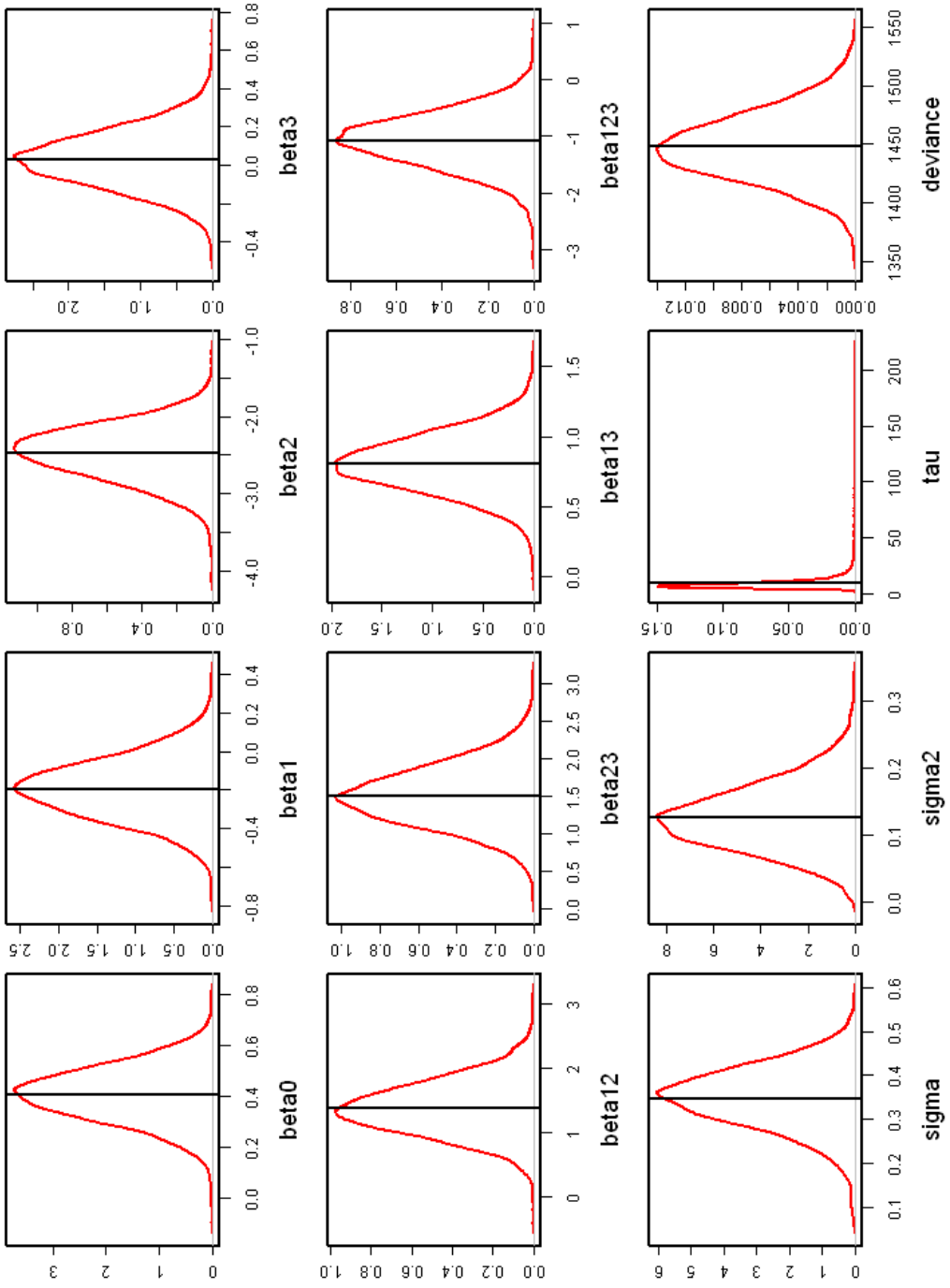

Figure A.3: Density plot for $B_\nu = 5$, $\tau_3 = 10^{-4}$, Thinning rate $= 40$

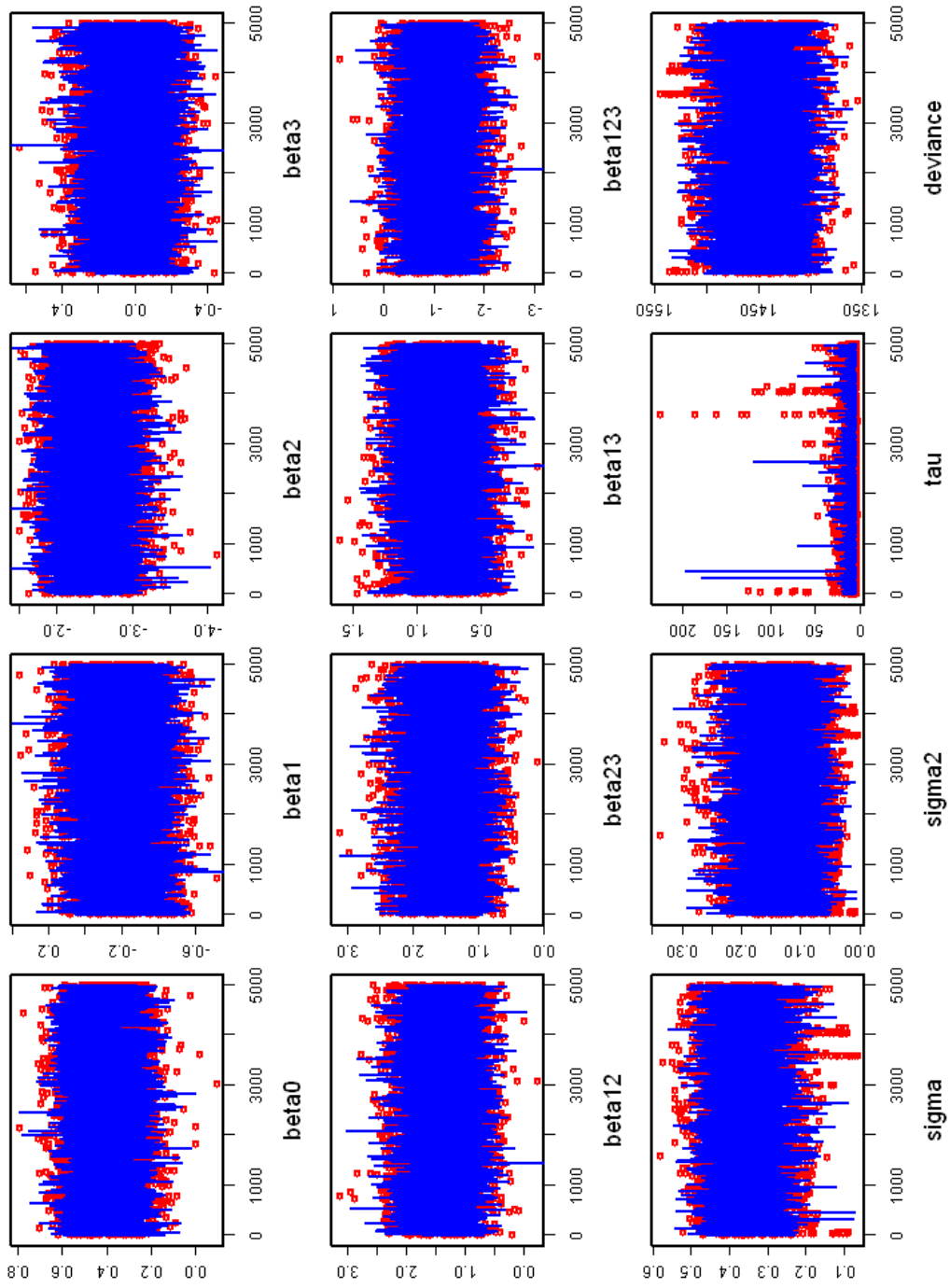

Figure A.4: Trace plot for $B_\nu = 5$, $\tau_3 = 10^{-4}$, Thinning rate $= 40$

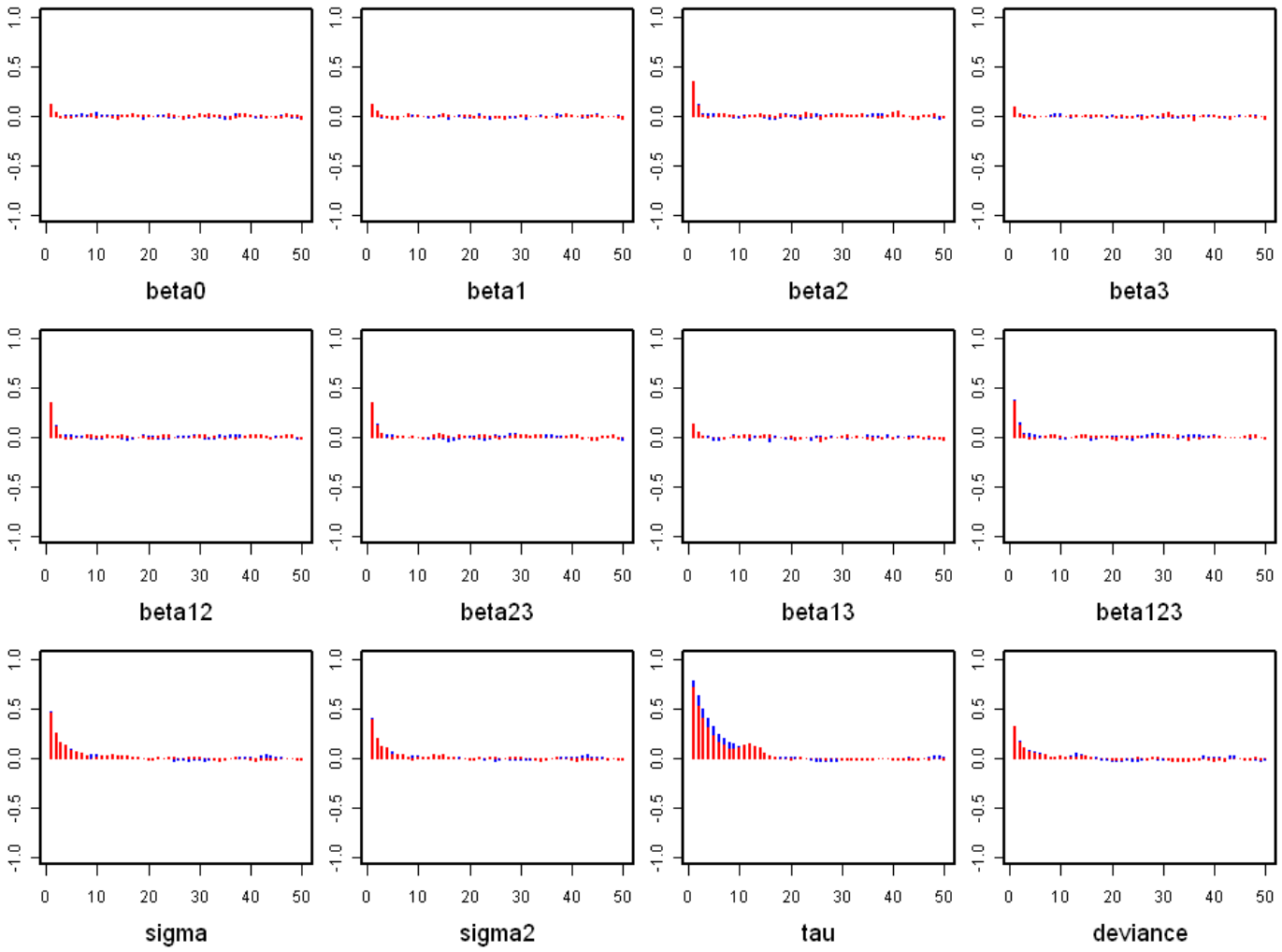

Figure A.5: Autocorrelation plot for $B_\nu = 5$, $\tau_3 = 10^{-4}$, Thinning rate = 40

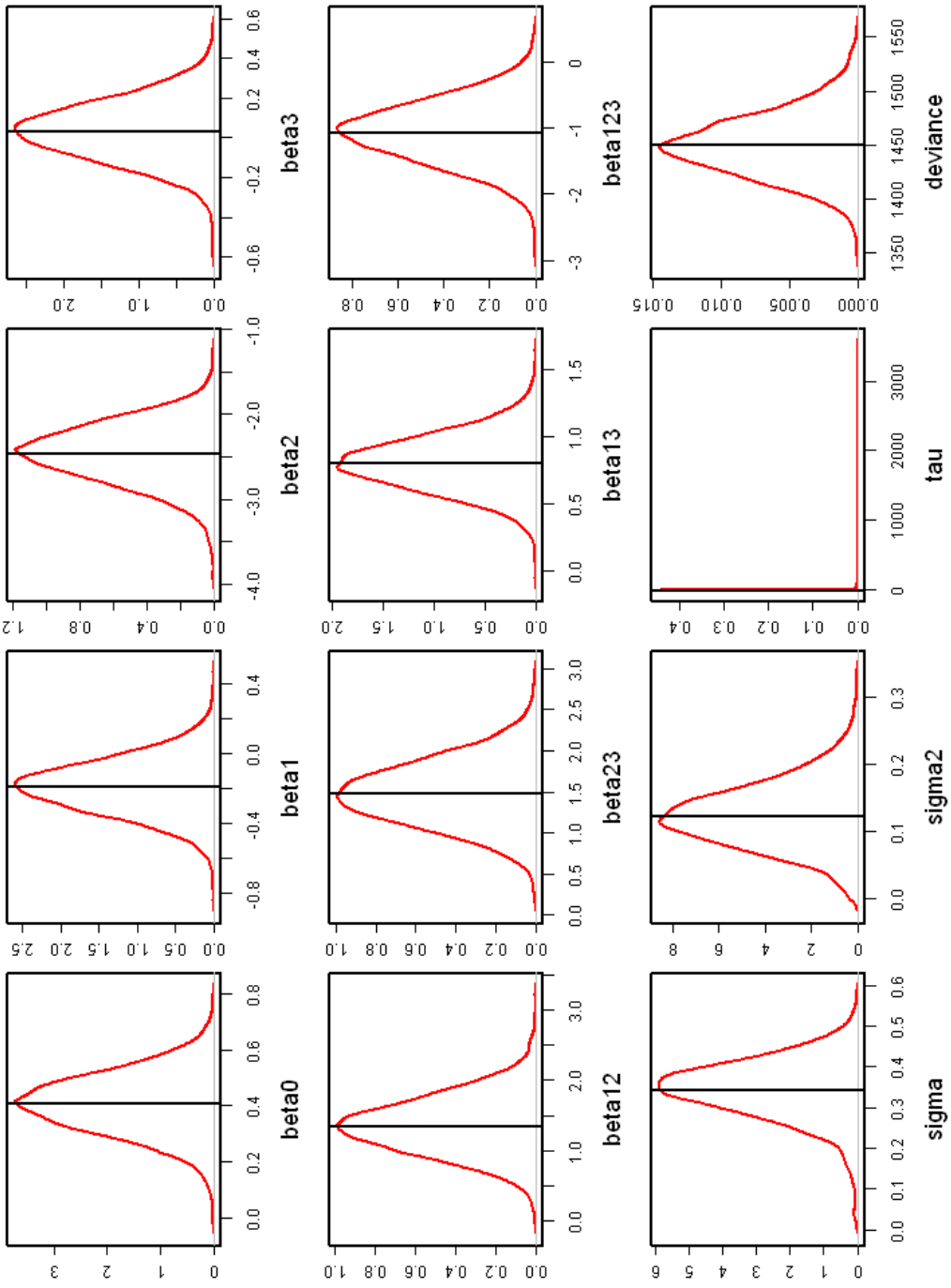

Figure A.6: Density plot for $B_\nu = 5$, $\tau_3 = 10^{-4}$, Thinning rate $= 25$

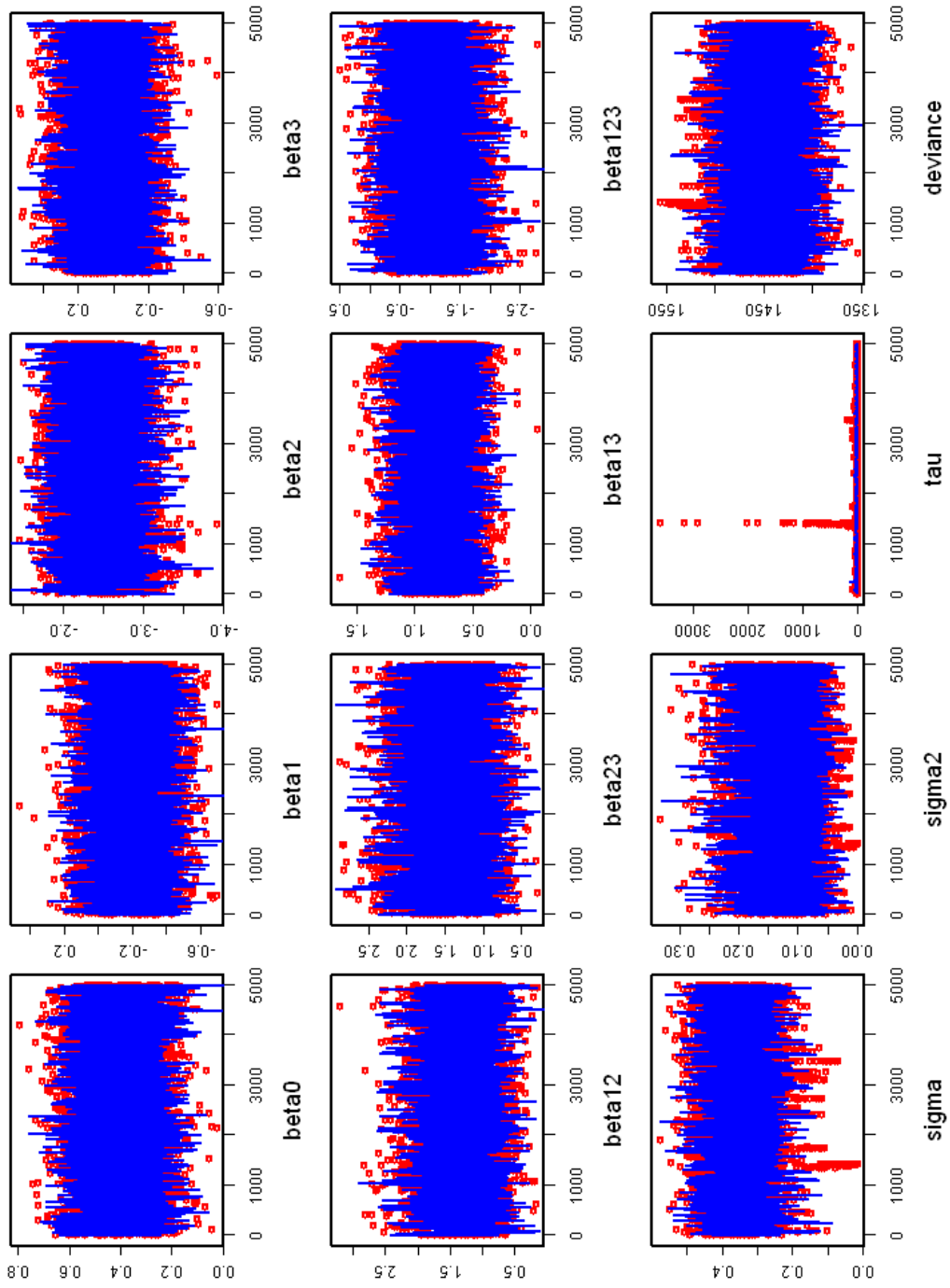

Figure A.7: Trace plot for $B_\nu = 5$, $\tau_3 = 10^{-4}$, Thinning rate $= 25$

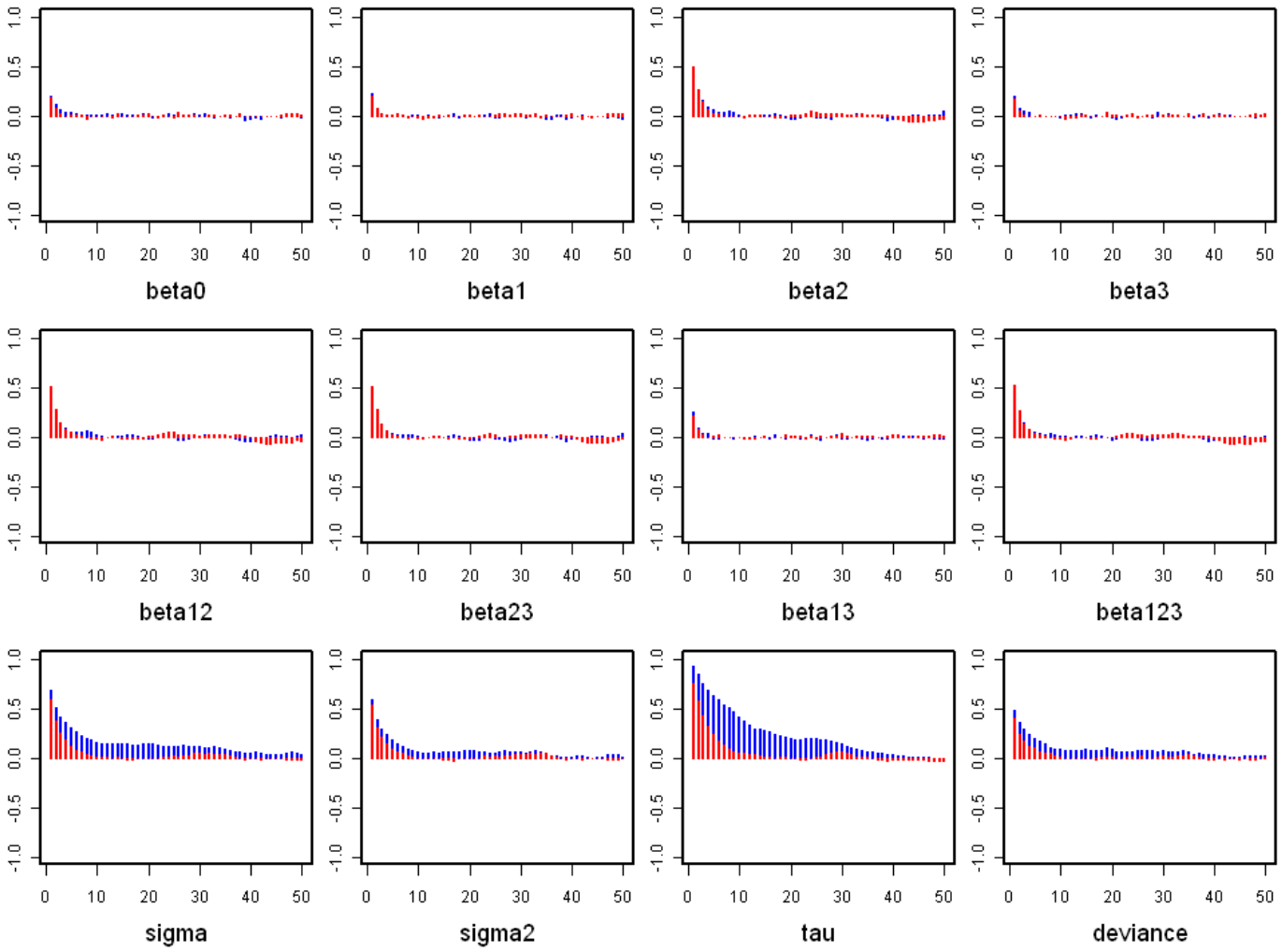

Figure A.8: Autocorrelation plot for $B_\nu = 5$, $\tau_3 = 10^{-4}$, Thinning rate = 25

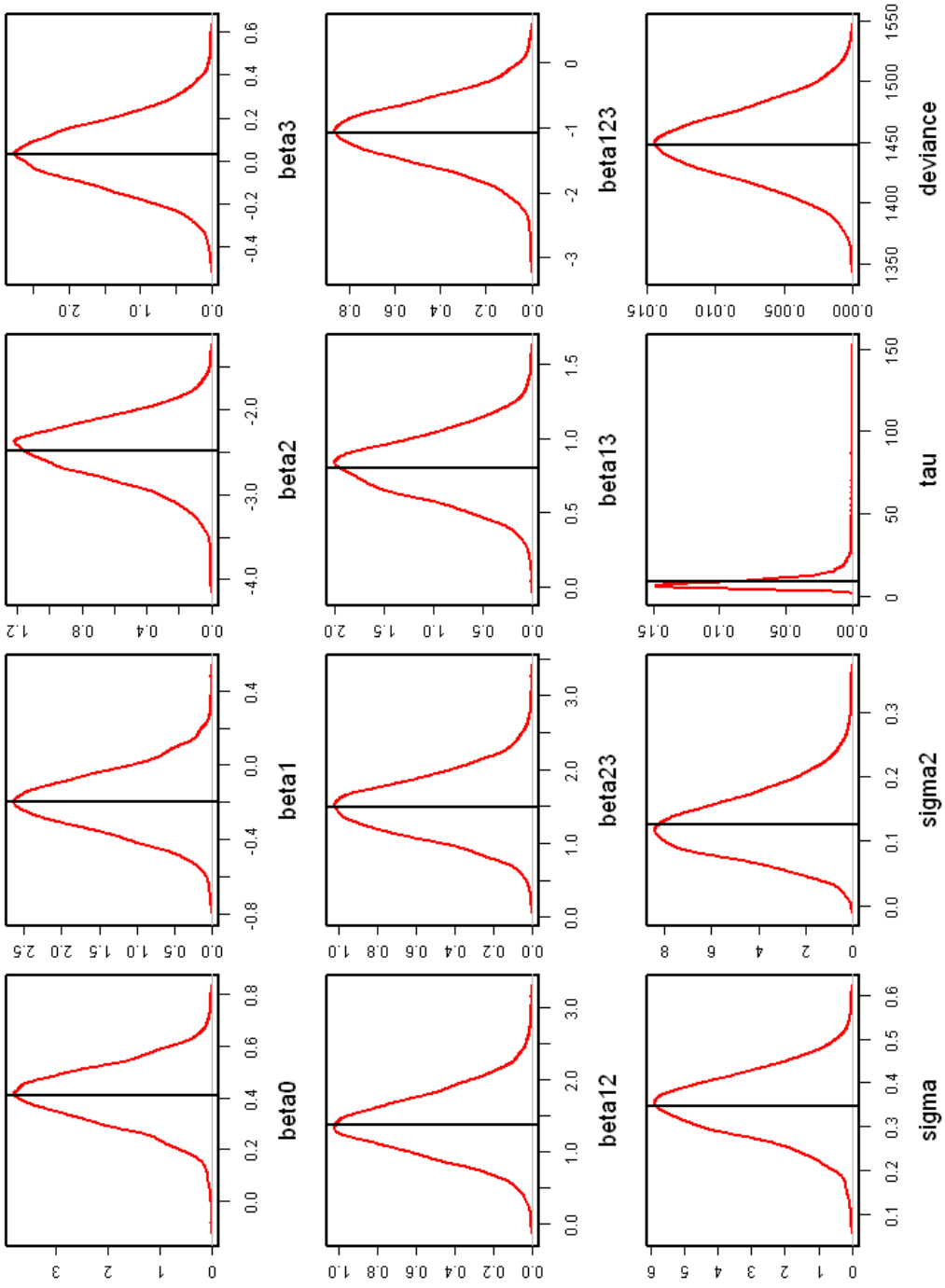

Figure A.9: Density plot for $B_\nu = 1$, $\tau_3 = 10^{-4}$, Thinning rate $= 25$

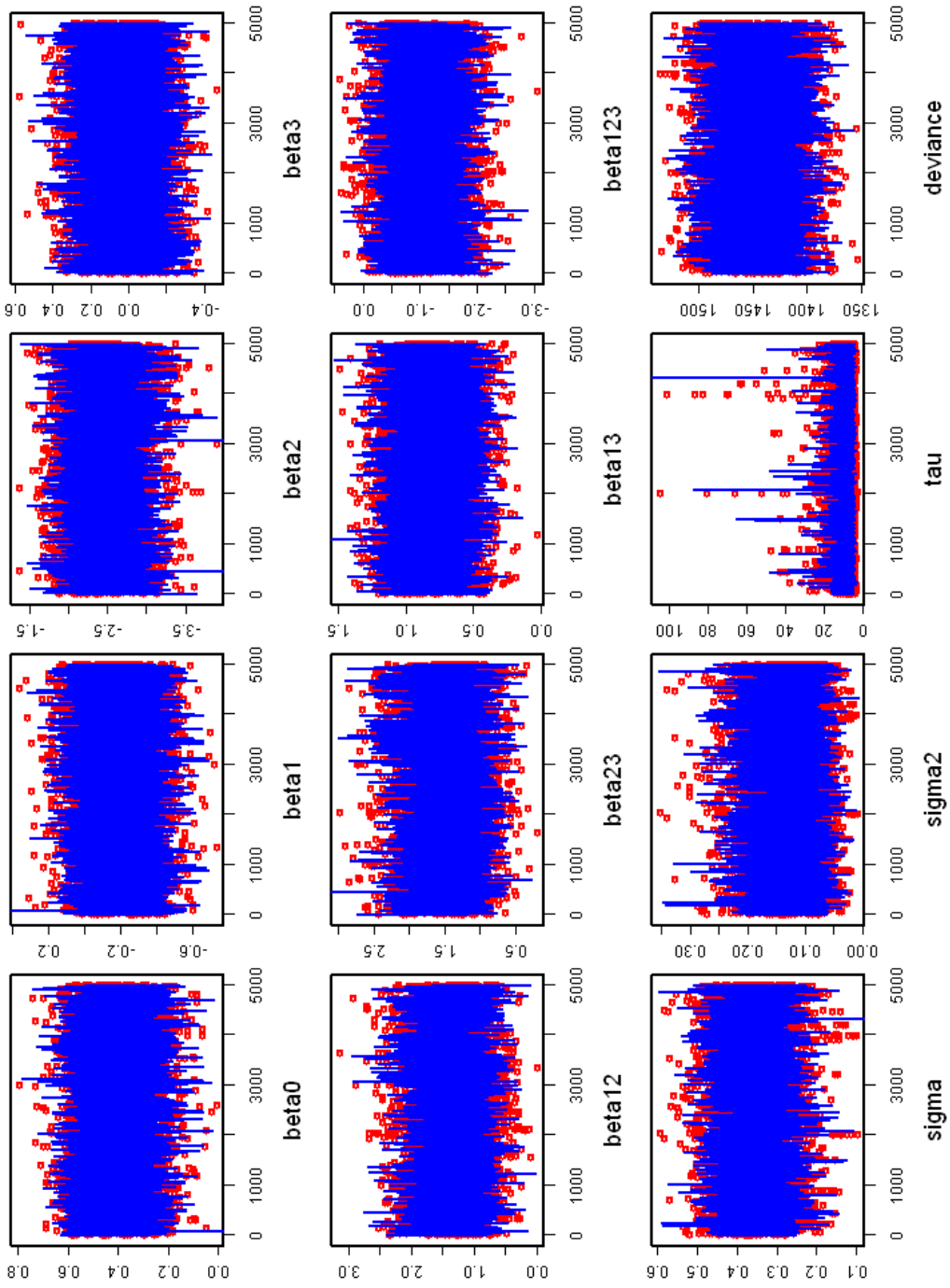

Figure A.10: Trace plot for $B_\nu = 1$, $\tau_3 = 10^{-4}$, Thinning rate $= 25$

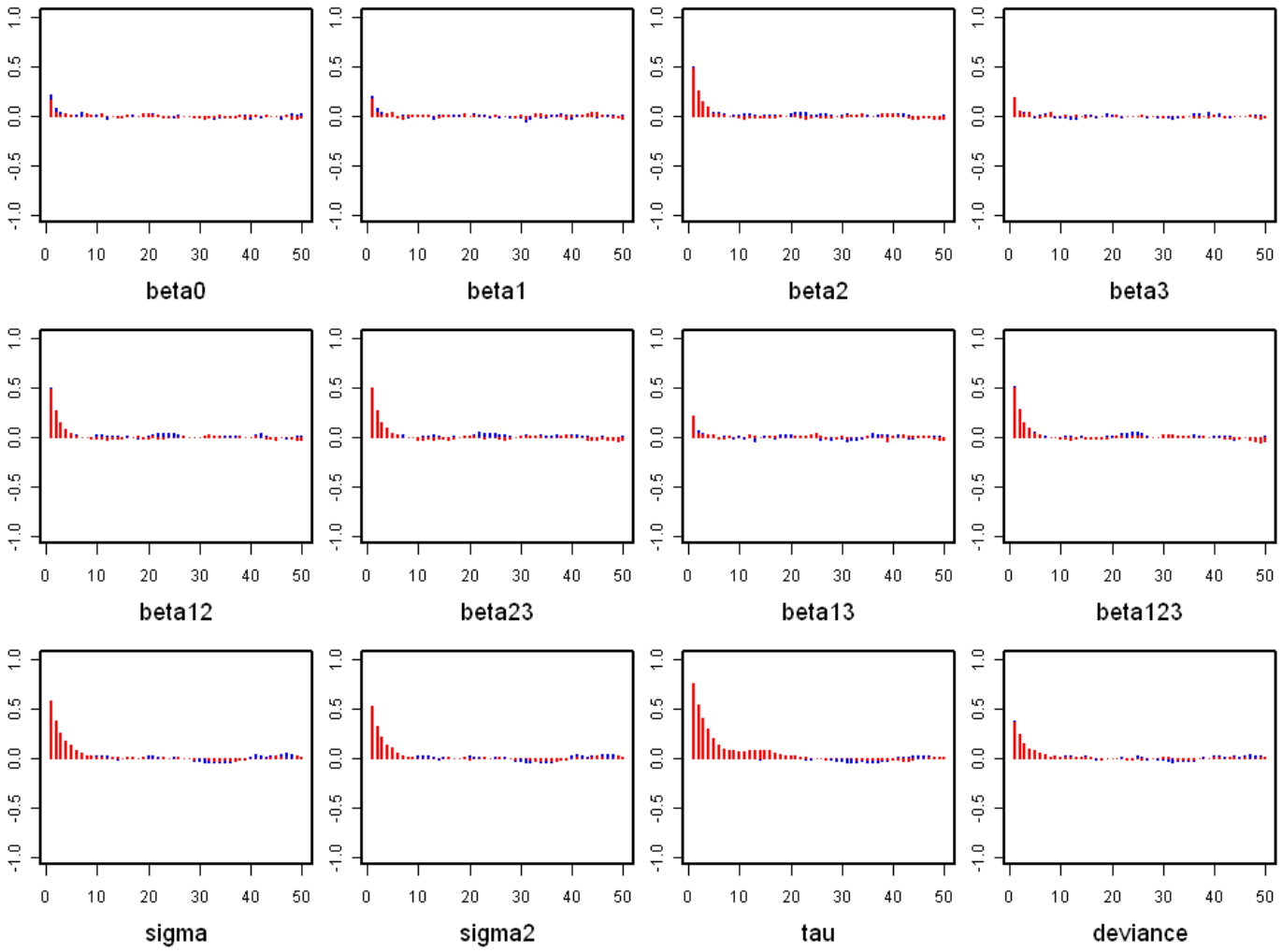

Figure A.11: Autocorrelation plot for $B_\nu = 1$, $\tau_3 = 10^{-4}$, Thinning rate = 25

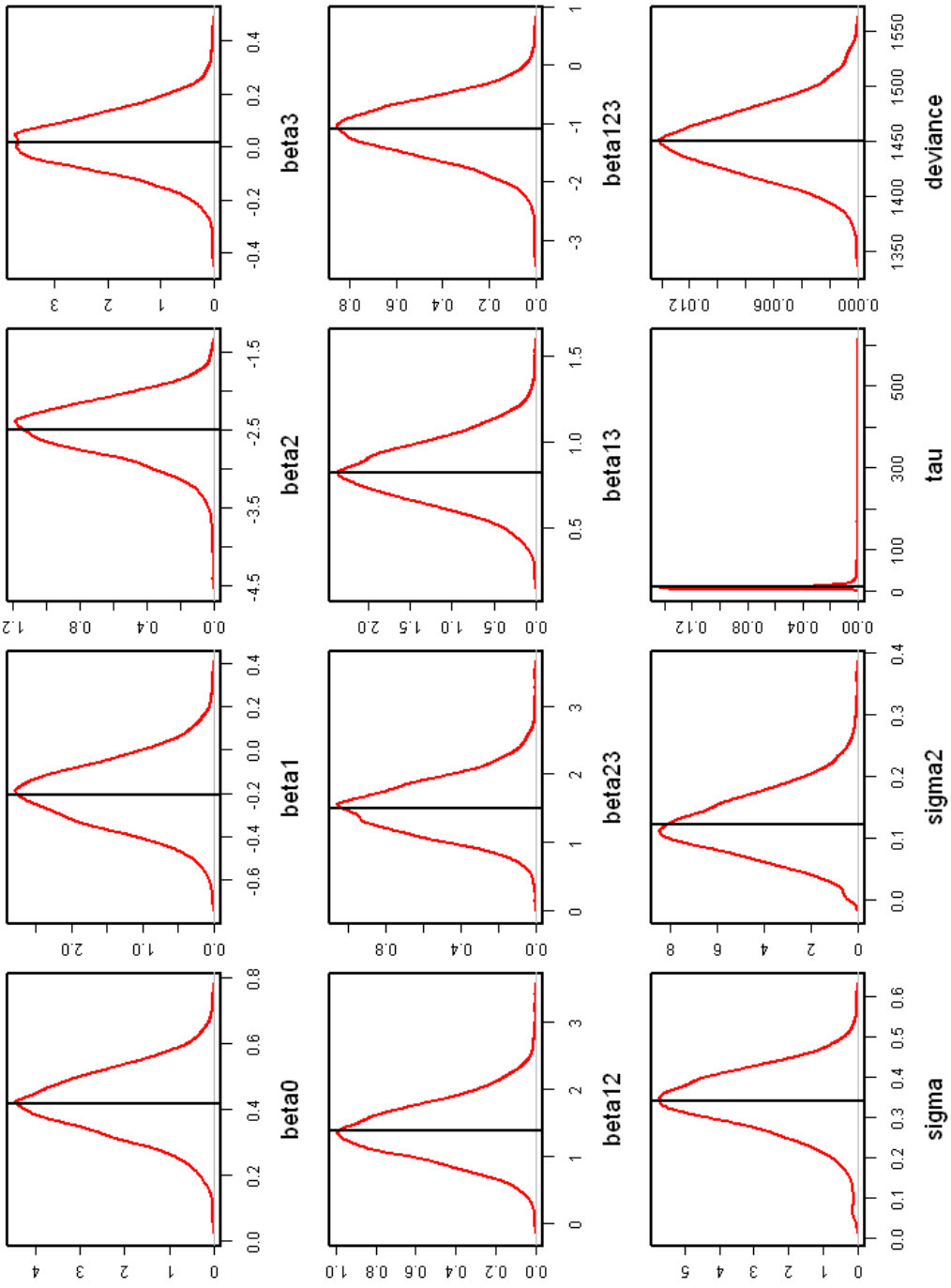

Figure A.12: Density plot for $B_\nu = 5$, $\tau_3 = 1/.144^2$, Thinning rate $= 25$

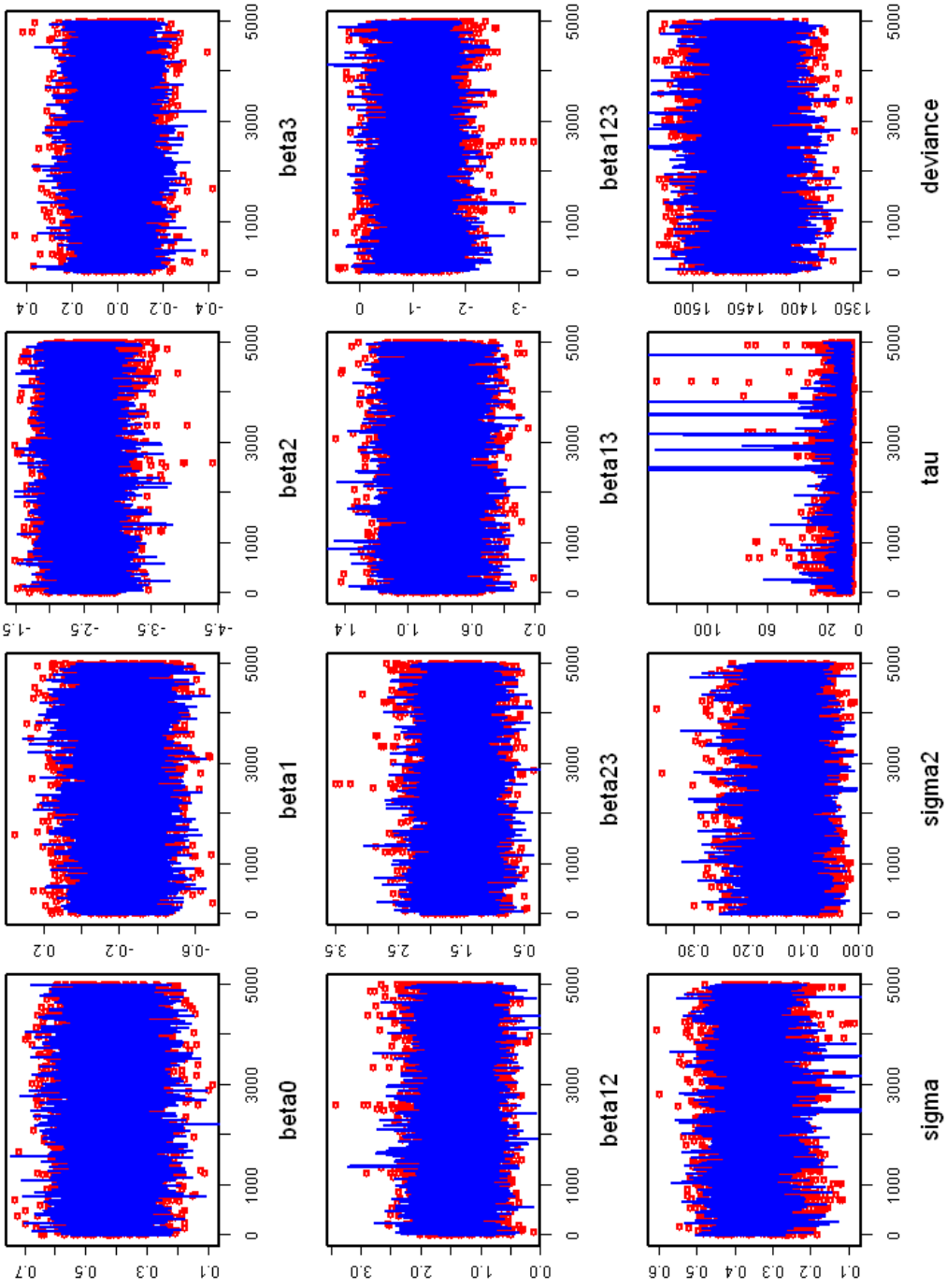

Figure A.13: Trace plot for $B_\nu = 5$, $\tau_3 = 1/.144^2$, Thinning rate $= 25$

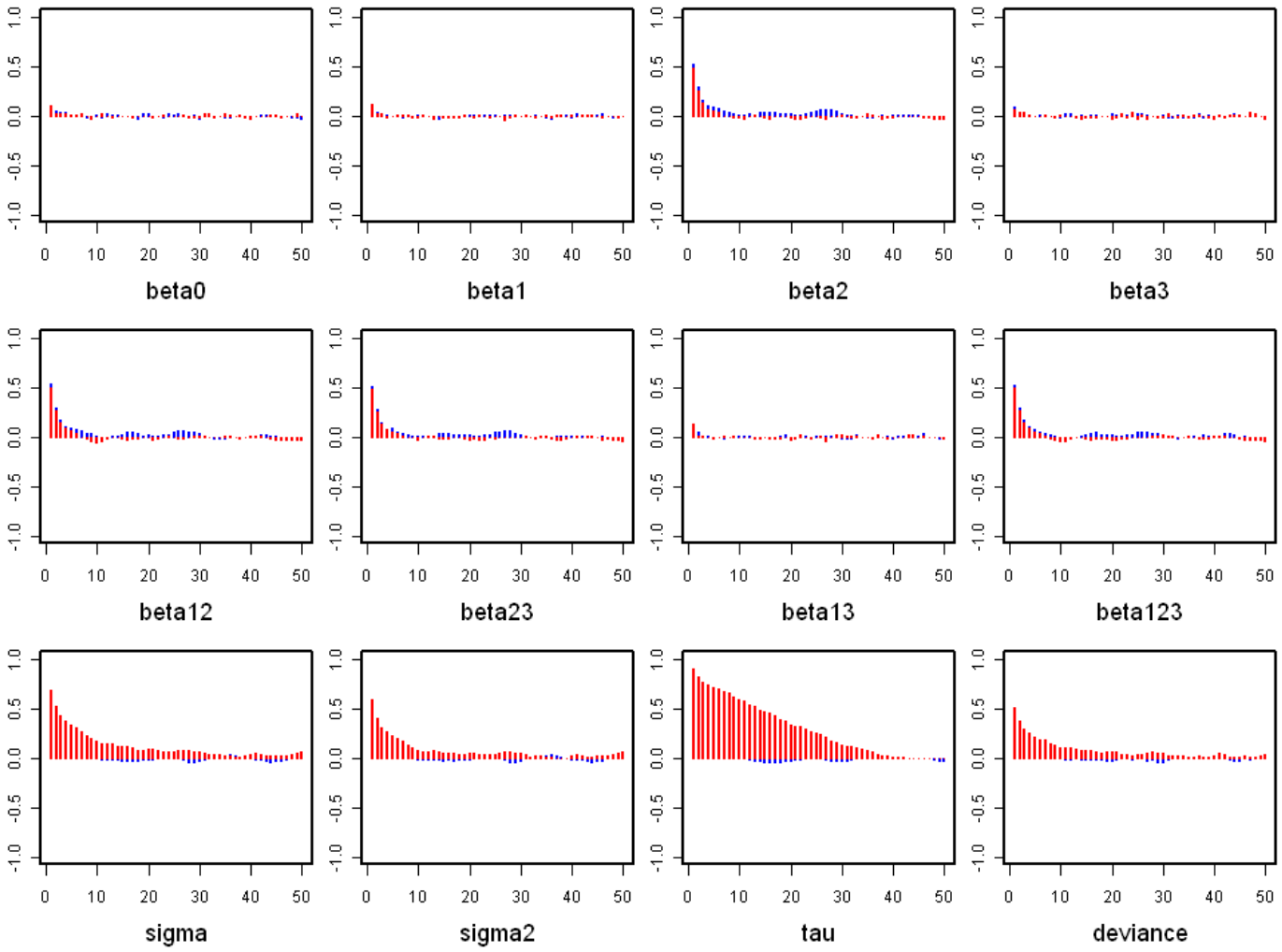

Figure A.14: Autocorrelation plot for $B_\nu = 5$, $\tau_3 = 1/.144^2$, Thinning rate = 25

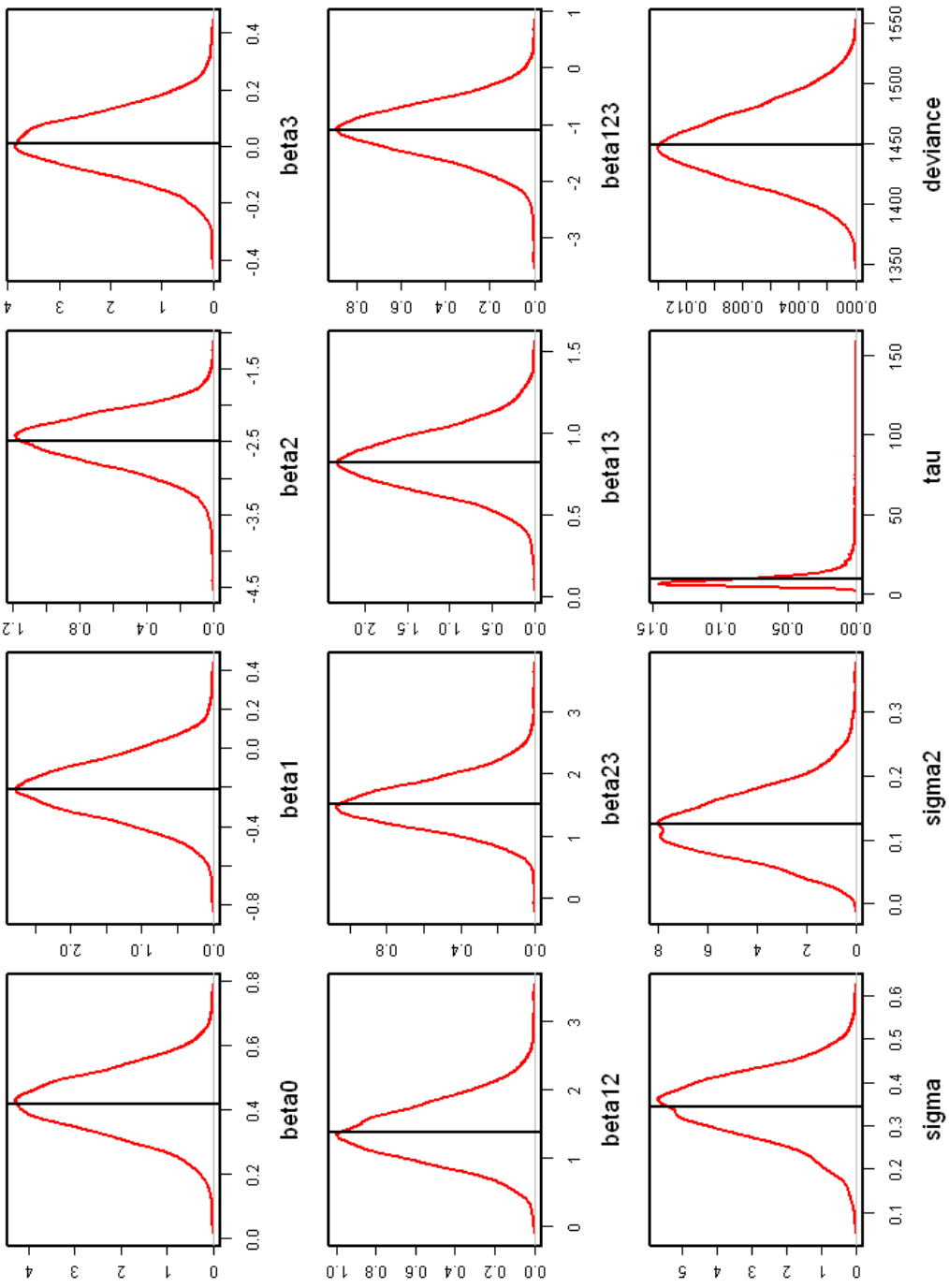

Figure A.15: Density plot for $B_\nu = 1$, $\tau_3 = 1/.144^2$, Thinning rate $= 25$

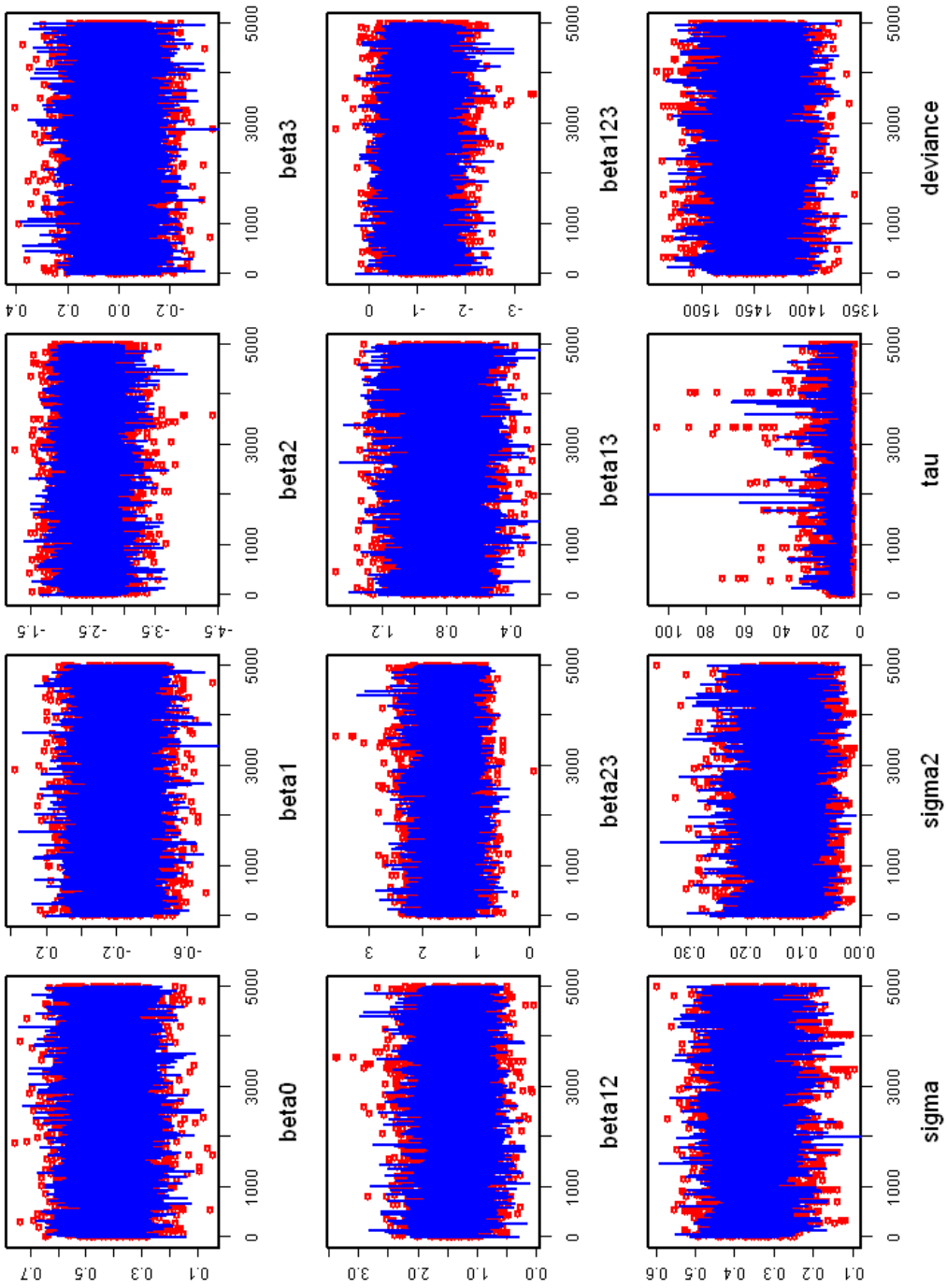

Figure A.16: Trace plot for $B_\nu = 1$, $\tau_3 = 1/.144^2$, Thinning rate $= 25$

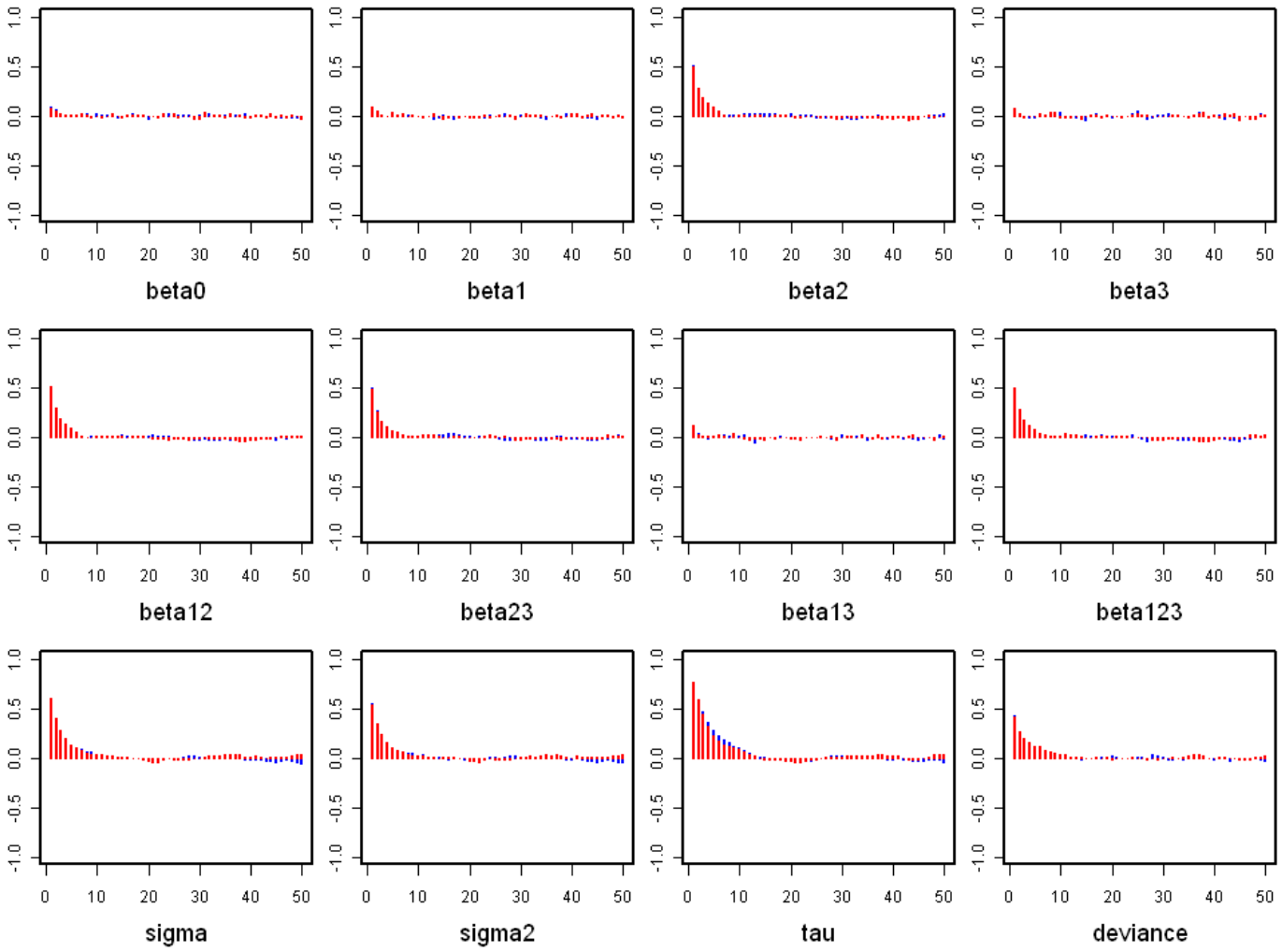

Figure A.17: Autocorrelation plot for $B_\nu = 1$, $\tau_3 = 1/.144^2$, Thinning rate = 25

A.3 Criteria for Positivity

Recall that the magnitude of the immune response is measured using an ELISPOT assay (Miyahira et al., 1995). In this procedure, cells extracted from vaccinated animals are exposed to HIV-1 fragments (peptides). The magnitude of the response is characterized by the number of cells releasing cytokines which encourage immune response. The number of cells in each well of the assay is preferably a million, but, since cells are at a premium in these studies, the number is typically around 100,000-200,000. The number of reactive cells is rescaled and reported as the number of spot-forming cells (SFC) per million.

For each animal, we obtain a collection of cells with no peptide stimulus and count the number of cells releasing cytokines. This background measurement is recorded twice, and the average of these two scores is reported as the background signal for this animal. Next, for each collection of cells, a single peptide stimulus is introduced, and the number of SFC is again recorded. This procedure is repeated for every peptide in the Gag and Nef genes. A collection of cells reacted positively to a peptide if the number of SFC observed after exposure to the peptide is at least 55 and at least four times the background for the animal whose cells are being tested.

Because each animal has a different background measurement, the peptides are judged according to different crtieria for positivity within each animal. The researchers prefer this method because it is standard in the field and is very conservative. Critics argue that this is too conservative, and that using individual backgrounds results in a loss of many positive reactives. That is, many peptides are misclassified as having a negative reaction when they should be counted as positive. One might consider developing a more uniform method of modeling the background. This could be done by constructing some overall summary background measure (for example, the mean or median plus 2 standard deviations of the individual background measurements).

1.4.1 Precise Count Model

```
model{
for(i in 1:n){
        log(lambda[i])<- beta[1] + beta[2]*x1[i] + beta[3]*x2[i] +
                beta[4]*x1x2[i]
        x[i]~dpois(lambda[i])
        }
## Normal priors on regression coefficients
for (g in 1:4) beta[g] ~ dnorm(0, 0.01)
}
```

1.4.2 Lower Limit Model

```
model{
for(i in 1:n){
        log(lambda[i])<- beta[1] + beta[2]*x1[i] + beta[3]*x2[i] +
                beta[4]*x1x2[i]
        j[i]~dpois(lambda[i])
        }
## Normal priors on regression coefficients
for (g in 1:4) beta[g] ~ dnorm(0, 0.01)
}
```

1.4.3 Interval Censored Poisson (ICP) Model

```
model{

C <- 1000000

for (i in 1:n) {
        zeros[i] <- 0
        zeros[i] ~ dpois( phi[i] )
        phi[i] <- -log(Lsum[i]) + C
        theta[i] <- exp( beta[1] + beta[2]*x1[i] + beta[3]*x2[i] +
                beta[4]*x1x2[i]  )

        for (t in 1:(w+1)) {
                L[i,t] <- step(k[i] - (j[i]+t-1)) *
                    exp(-theta[i])*pow(theta[i],  j[i]+t-1)/
                        exp(loggam( j[i]+t-1+1))
                                }
                Lsum[i] <- sum(L[i,])
       }
## Normal priors on regression coefficients
for (g in 1:4) beta[g] ~ dnorm(0, 0.01)
 }
```

1.4.4 Interval Censored Poisson with Zero Inflation Adjustment (ICP ZIP) Model

```
model{

C <- 1000000
for (i in 1:n) {
        zeros[i] <- 0
        zeros[i] ~ dpois( phi[i] )
        phi[i] <- -log(Lz[i]) + C
        theta[i] <- exp( beta[1] + beta[2]*x1[i] + beta[3]*x2[i] +
                beta[4]*x1x2[i]  )

        for (t in 1:(w+1)) {
                                ## Likelihood
                L[i,t] <- step(k[i] - (j[i]+t-1)) *
                        exp(-theta[i])*pow(theta[i],  j[i]+t-1)/
                                exp(loggam( j[i]+t-1+1))
                                }
                Lsum[i] <- sum(L[i,])
                Lz[i] <-  p0[i] * equals( j[i], 0 ) + (1-p0[i]) * Lsum[i]

        p0[i] ~ dbeta(1, 1) # Independent uniform priors for p0
       }

## Normal priors on regression coefficients
for (g in 1:4) beta[g] ~ dnorm(0, 0.01)
 }
```

A.5 ICP Regression Results for $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$

In the following section, we present the fixed and mixed effect model results for data generated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$.

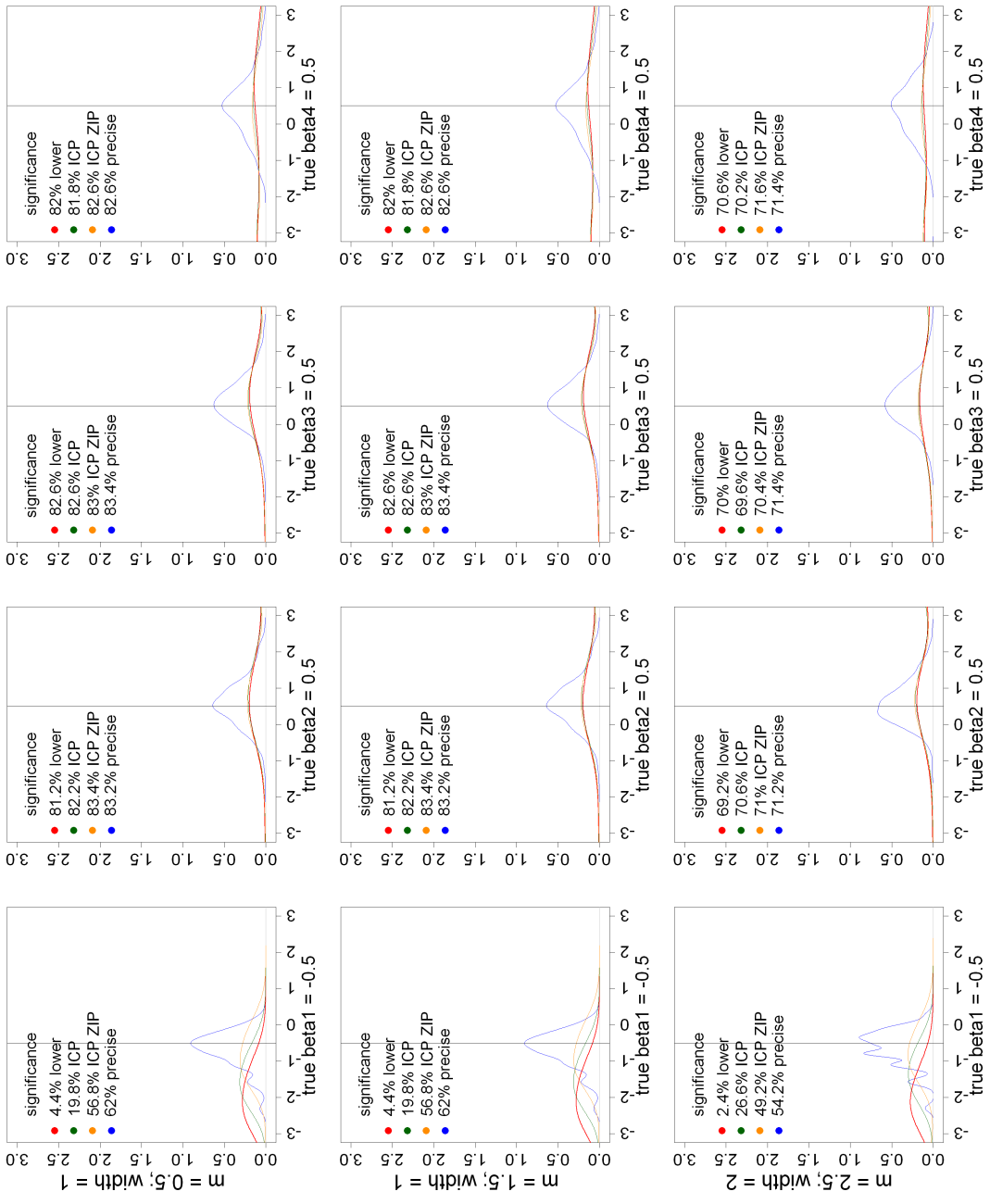

Figure A.18: Posterior medians for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

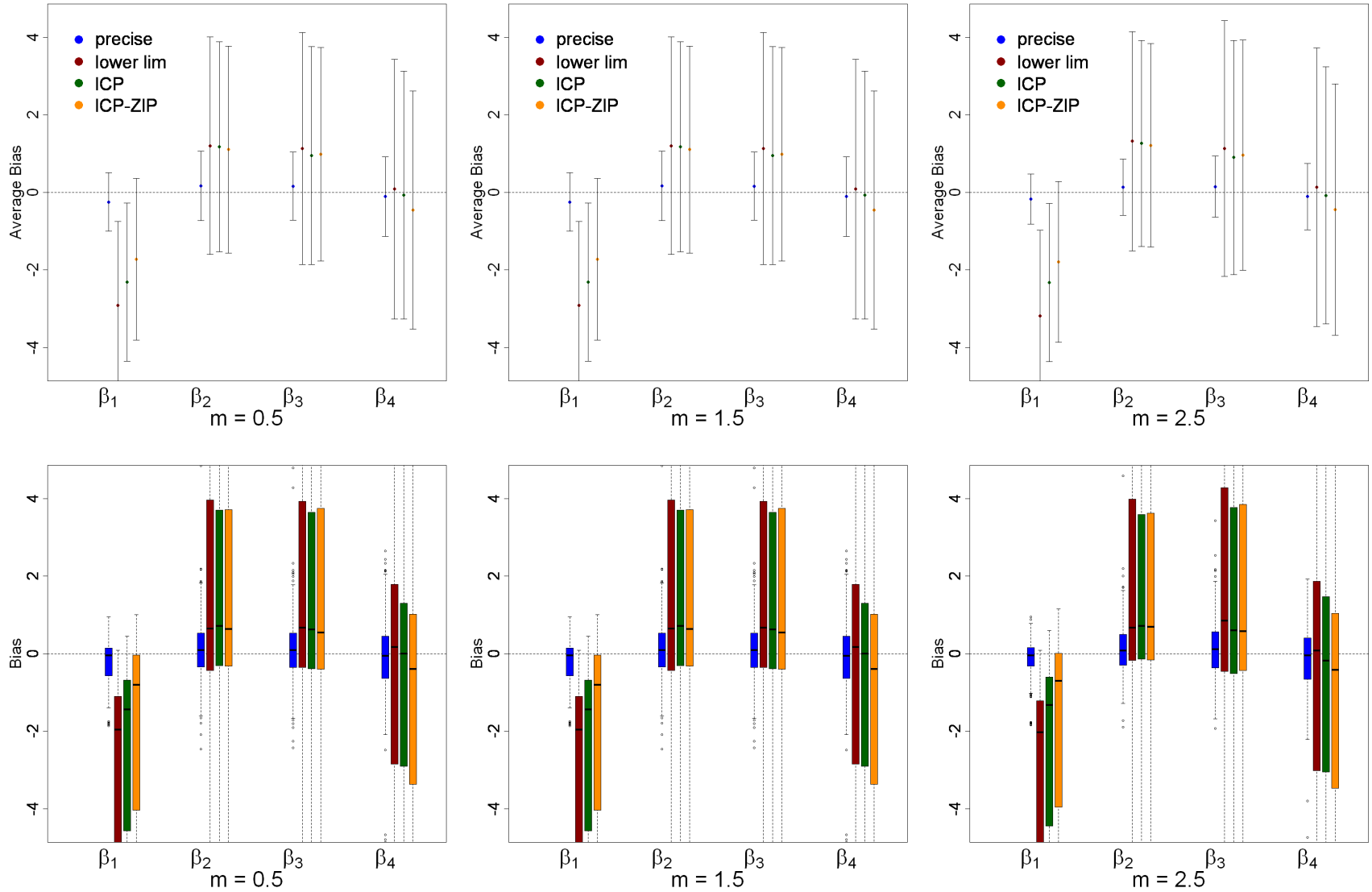

Figure A.19: Bias for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

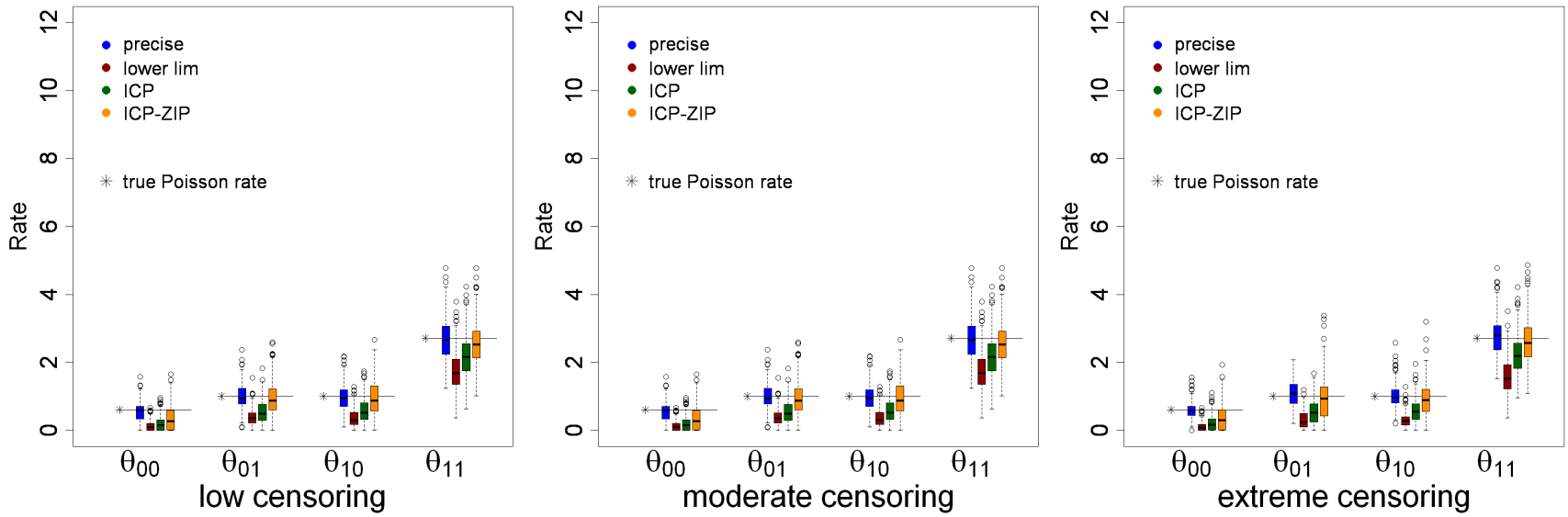

Figure A.20: Poisson rates for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

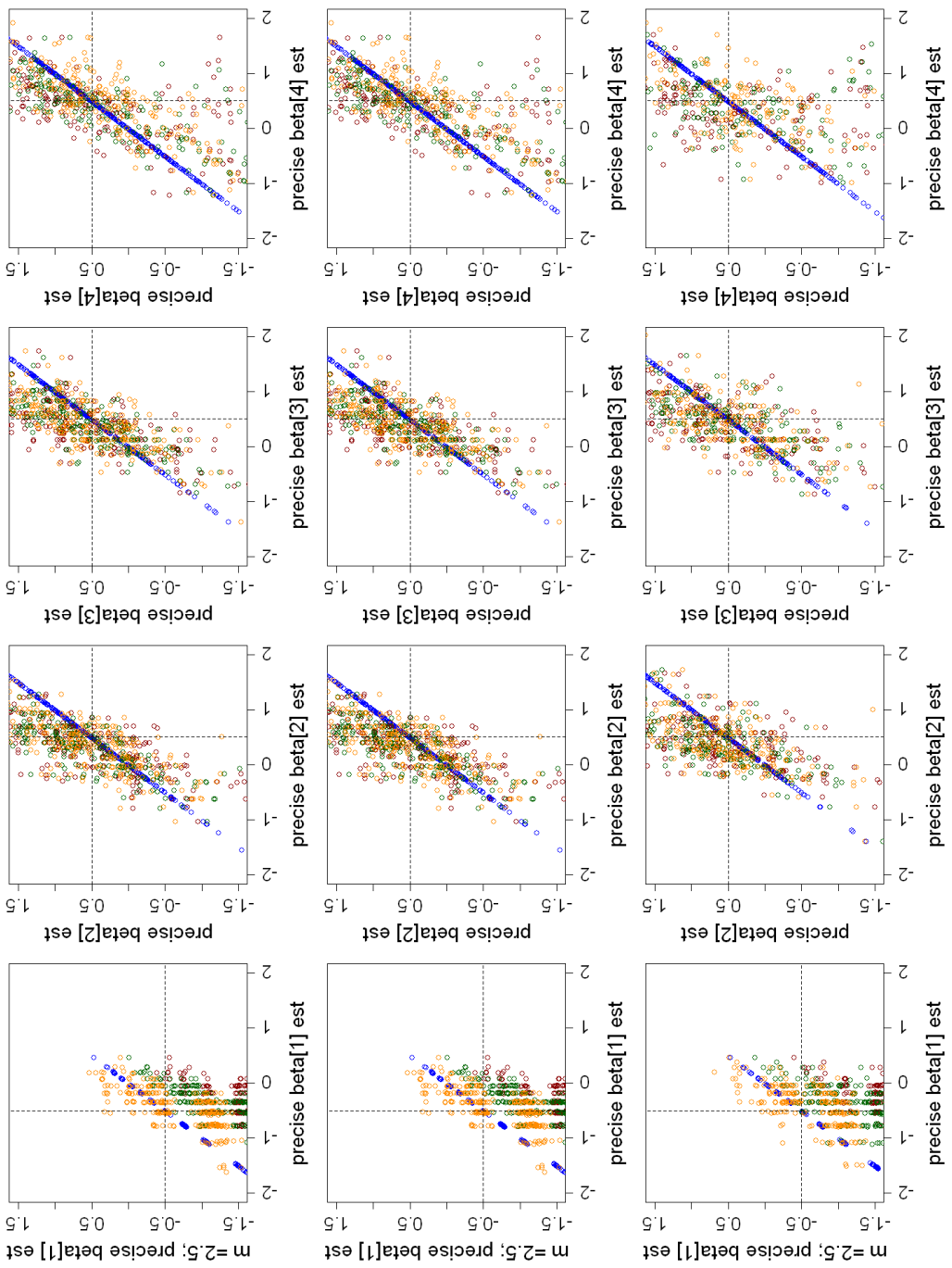

Figure A.21: Lower limit and interval model estimates plotted against precise count model estimates for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

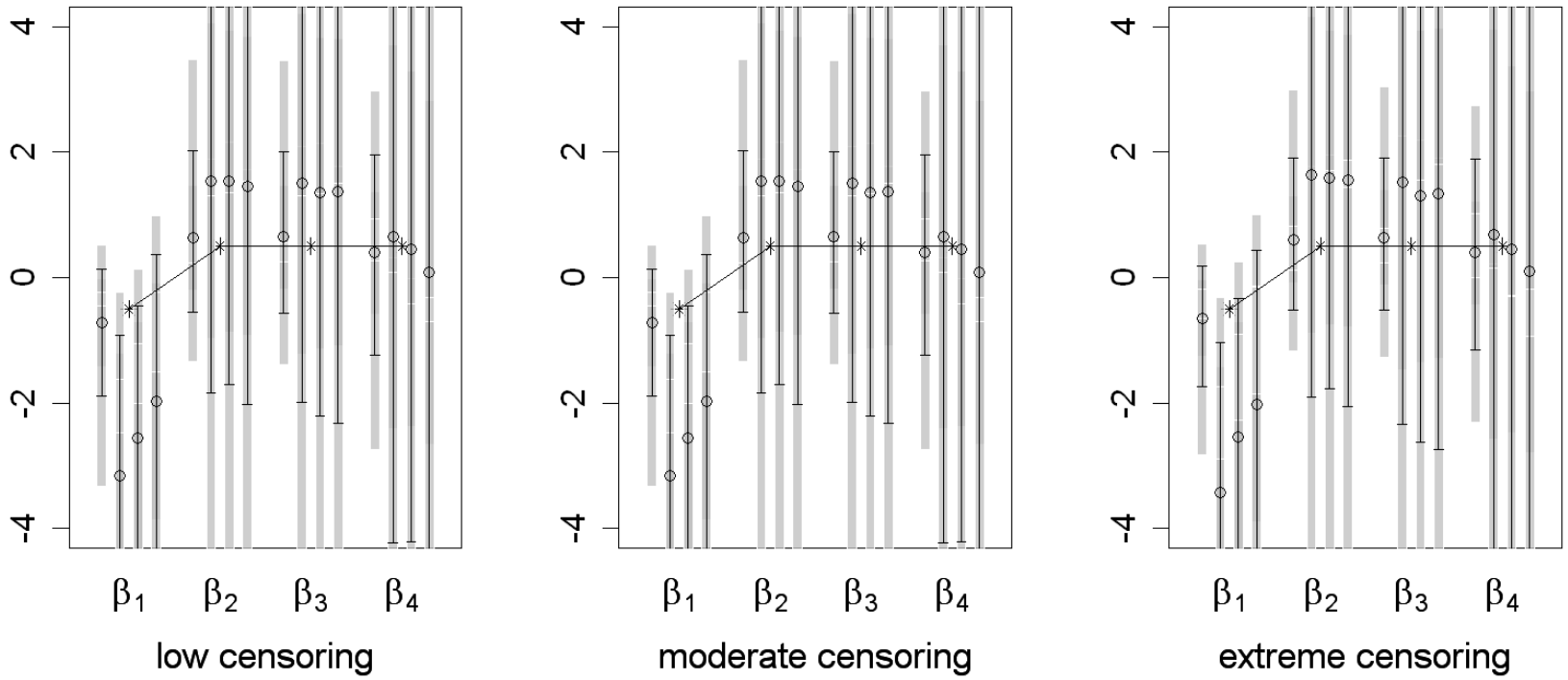

Figure A.22: Summary of the coverage of 95% credible intervals for regression coefficient estimates for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in fixed effect model

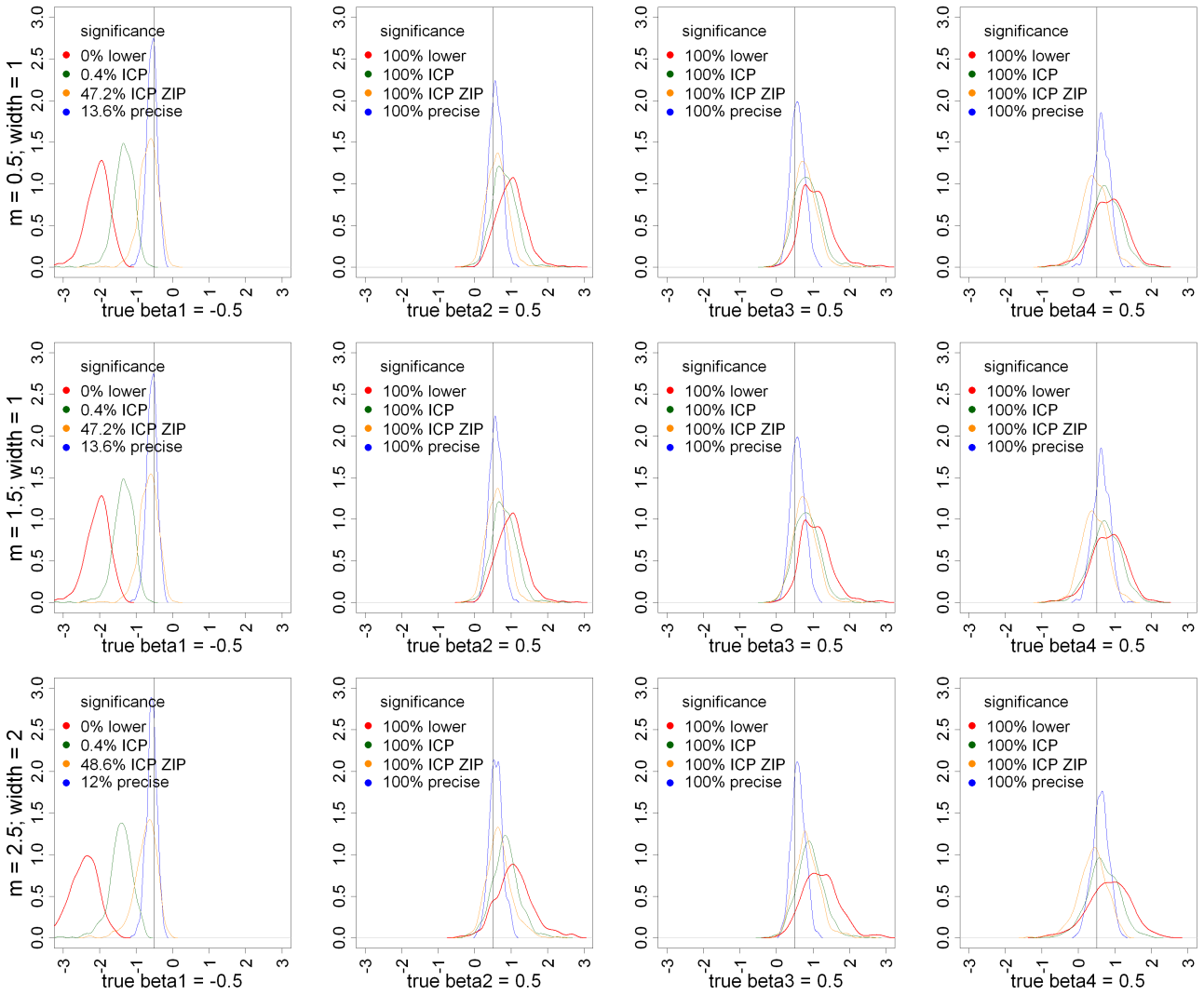

Figure A.23: Posterior medians for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in mixed effect model

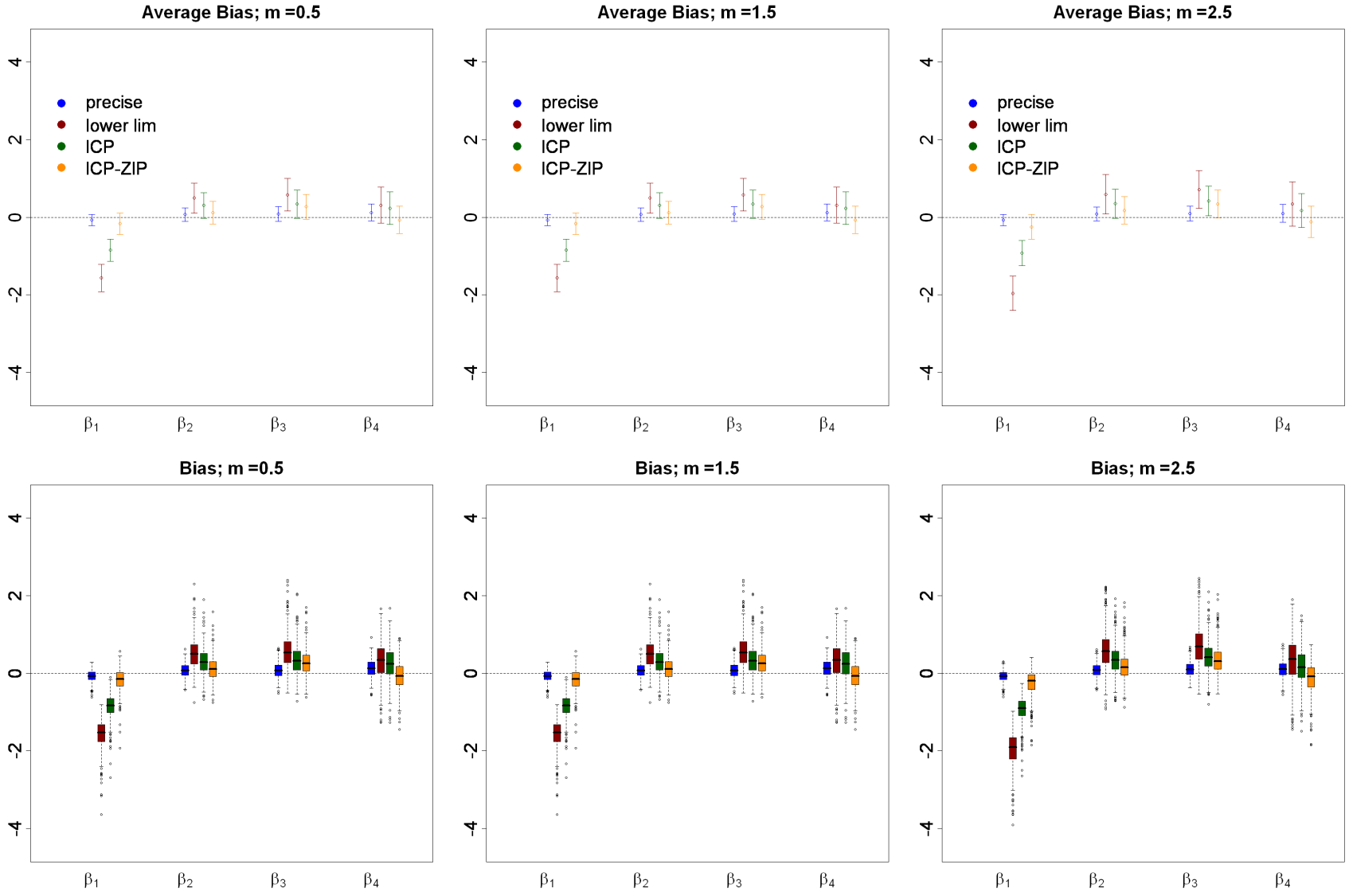

Figure A.24: Bias for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in mixed effect model

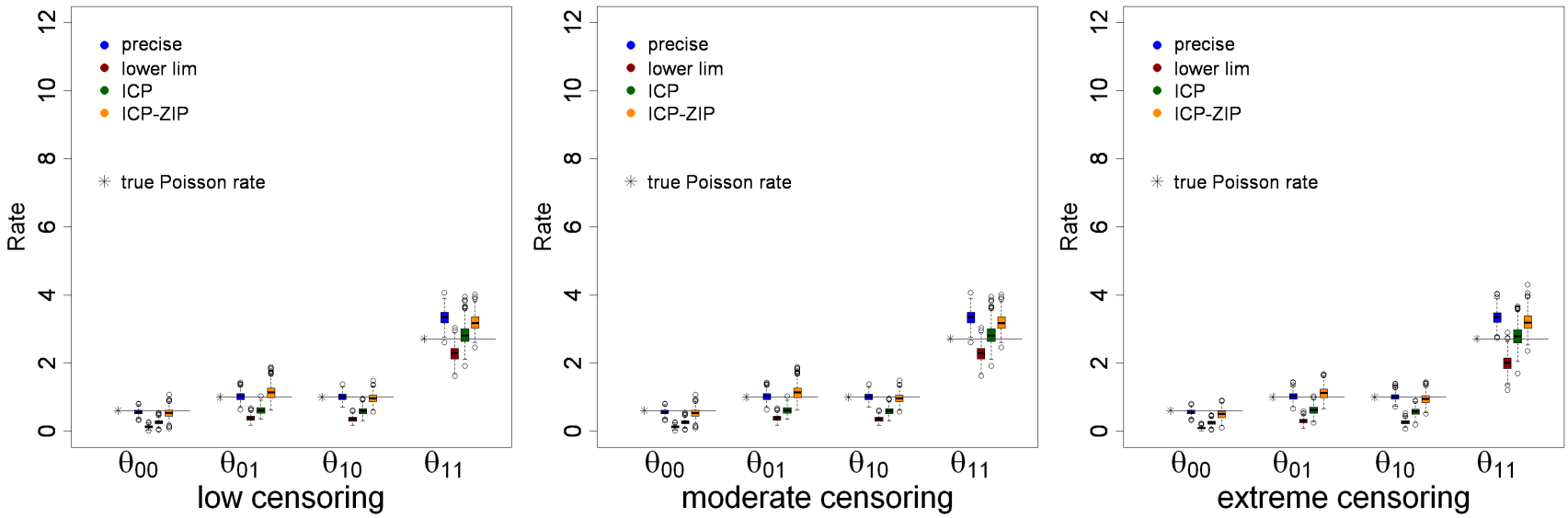

Figure A.25: Poisson rates for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in mixed effect model

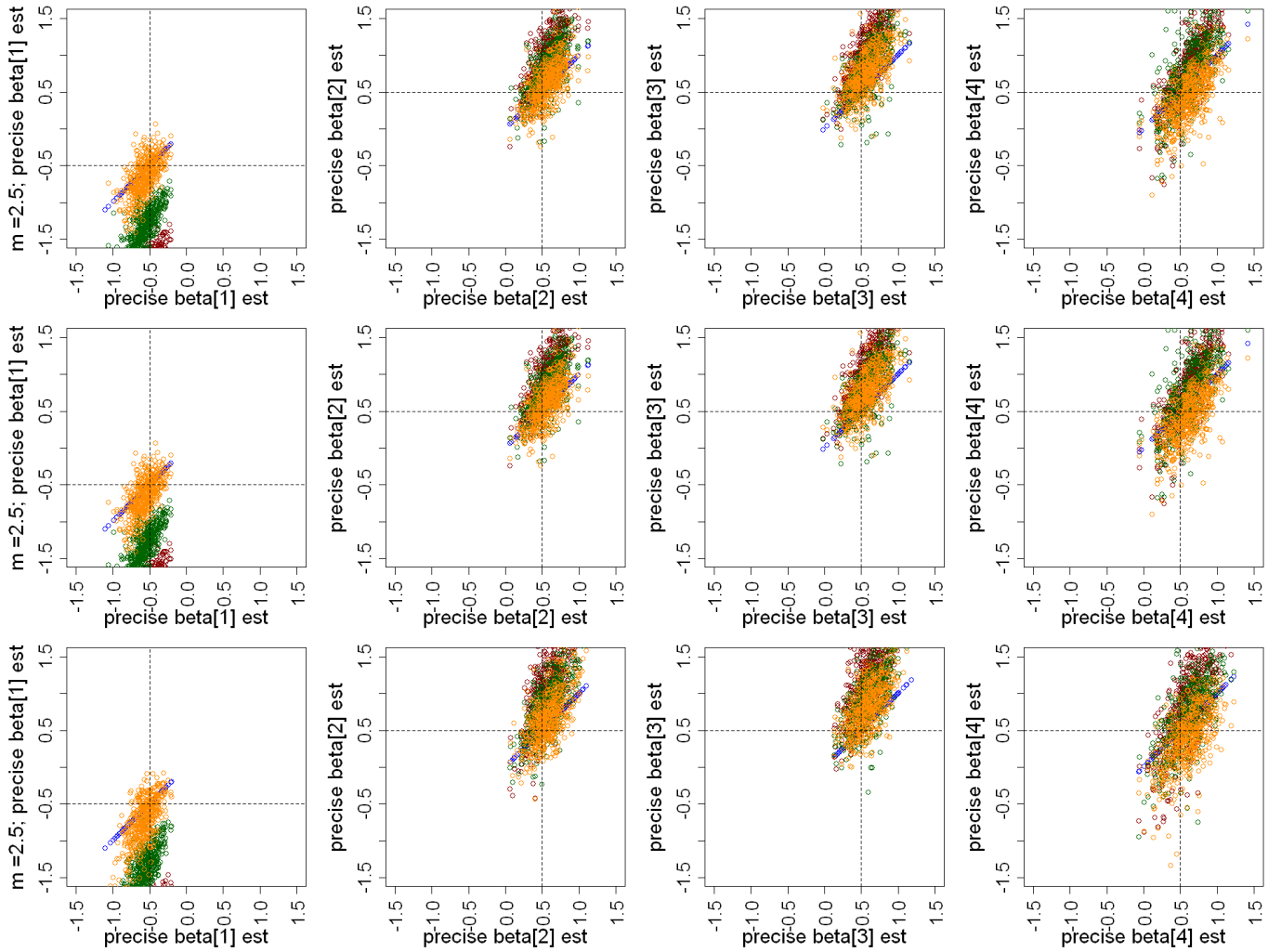

Figure A.26: Lower limit and interval model estimates plotted against precise count model estimates for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in mixed effect model

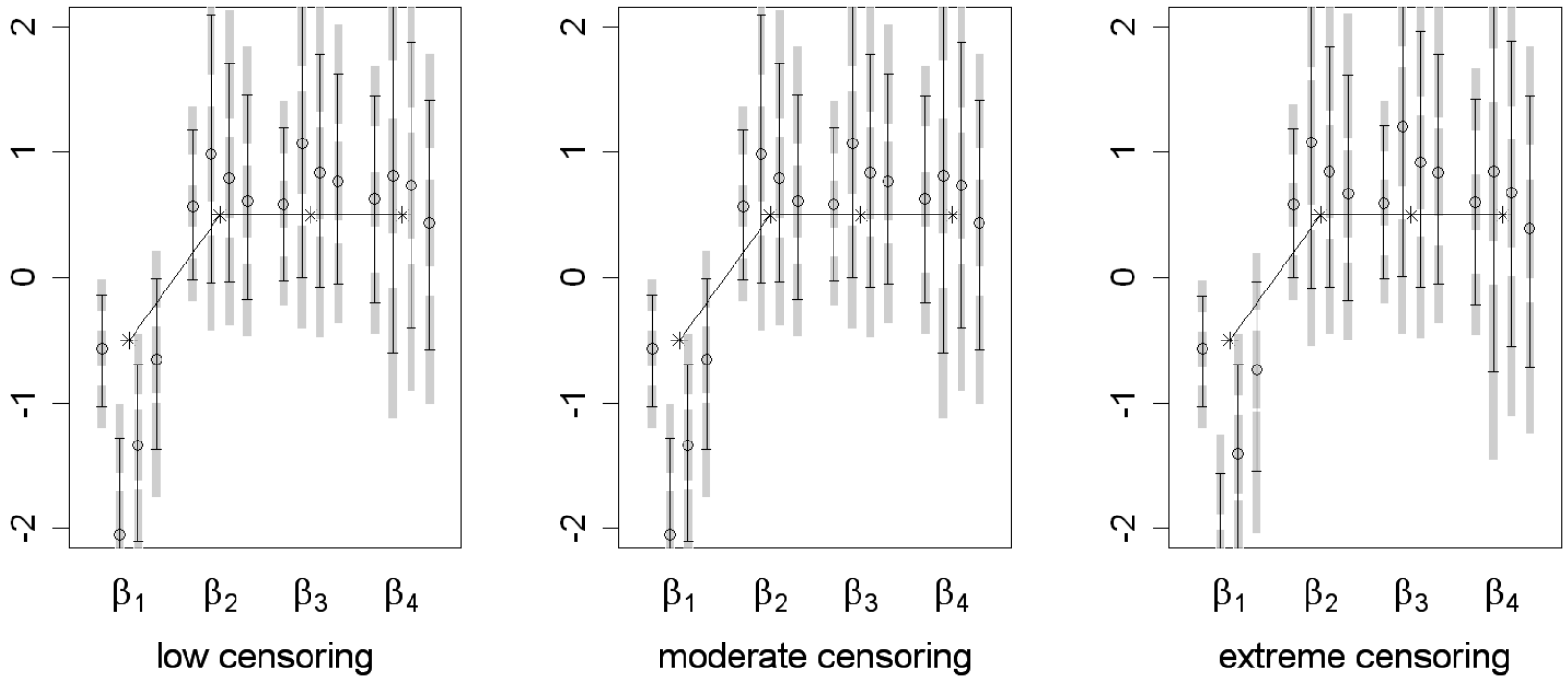

Figure A.27: Summary of the coverage of 95% credible intervals for regression coefficient estimates for data simulated from $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$ in mixed effect model

Table A.1: Coverage for lower limit and precise count fixed effects models and true $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$

Model	term	m	truth	avg.width	sd.width	coverage
Lower limit	β_1	0.50	-0.50	6.31	4.11	0.28
	β_2	0.50	0.50	7.77	4.42	0.67
	β_3	0.50	0.50	7.99	4.60	0.68
	β_4	0.50	0.50	9.34	4.68	0.69
	β_1	1.50	-0.50	6.31	4.11	0.28
	β_2	1.50	0.50	7.77	4.42	0.67
	β_3	1.50	0.50	7.99	4.60	0.68
	β_4	1.50	0.50	9.34	4.68	0.69
	β_1	2.50	-0.50	6.70	4.15	0.22
	β_2	2.50	0.50	8.19	4.26	0.58
	β_3	2.50	0.50	8.65	4.62	0.55
	β_4	2.50	0.50	9.96	4.51	0.58
Precise count	β_1	0.50	-0.50	2.03	1.10	0.79
	β_2	0.50	0.50	2.57	1.04	0.76
	β_3	0.50	0.50	2.58	1.11	0.77
	β_4	0.50	0.50	3.20	1.02	0.74
	β_1	1.50	-0.50	2.03	1.10	0.79
	β_2	1.50	0.50	2.57	1.04	0.76
	β_3	1.50	0.50	2.58	1.11	0.77
	β_4	1.50	0.50	3.20	1.02	0.74
	β_1	2.50	-0.50	1.91	0.76	0.67
	β_2	2.50	0.50	2.43	0.72	0.68
	β_3	2.50	0.50	2.43	0.80	0.64
	β_4	2.50	0.50	3.03	0.75	0.67

Table A.2: Coverage for ICP fixed effects models and true $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$

Model	term	m	truth	avg.width	sd.width	coverage
ICP	β_1	0.50	-0.50	6.13	4.03	0.53
	β_2	0.50	0.50	7.53	4.48	0.68
	β_3	0.50	0.50	7.91	4.66	0.69
	β_4	0.50	0.50	9.13	4.88	0.73
	β_1	1.50	-0.50	6.13	4.03	0.53
	β_2	1.50	0.50	7.53	4.48	0.68
	β_3	1.50	0.50	7.91	4.66	0.69
	β_4	1.50	0.50	9.13	4.88	0.73
	β_1	2.50	-0.50	6.50	4.03	0.46
	β_2	2.50	0.50	7.96	4.35	0.61
	β_3	2.50	0.50	8.61	4.73	0.60
	β_4	2.50	0.50	9.85	4.80	0.62
ICP ZIP	β_1	0.50	-0.50	6.57	4.11	0.76
	β_2	0.50	0.50	7.99	4.49	0.75
	β_3	0.50	0.50	8.26	4.67	0.77
	β_4	0.50	0.50	9.56	4.82	0.78
	β_1	1.50	-0.50	6.57	4.11	0.76
	β_2	1.50	0.50	7.99	4.49	0.75
	β_3	1.50	0.50	8.26	4.67	0.77
	β_4	1.50	0.50	9.56	4.82	0.78
	β_1	2.50	-0.50	6.97	4.13	0.68
	β_2	2.50	0.50	8.41	4.33	0.67
	β_3	2.50	0.50	9.02	4.74	0.66
	β_4	2.50	0.50	10.28	4.70	0.66

Table A.3: Coverage for lower limit and precise count mixed effects model and true $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$

Model	Term	m	Truth	Avg.width	SD.width	Coverage
Lower limit	β_1	0.50	-0.50	1.68	0.26	0.00
	β_2	0.50	0.50	2.13	0.30	0.93
	β_3	0.50	0.50	2.19	0.30	0.89
	β_4	0.50	0.50	2.84	0.41	0.99
	β_1	1.50	-0.50	1.68	0.26	0.00
	β_2	1.50	0.50	2.13	0.30	0.93
	β_3	1.50	0.50	2.19	0.30	0.89
	β_4	1.50	0.50	2.84	0.41	0.99
	β_1	2.50	-0.50	1.95	0.39	0.00
	β_2	2.50	0.50	2.43	0.41	0.90
	β_3	2.50	0.50	2.48	0.42	0.86
	β_4	2.50	0.50	3.18	0.52	0.98
Precise count	β_1	0.50	-0.50	0.89	0.10	0.99
	β_2	0.50	0.50	1.19	0.13	1.00
	β_3	0.50	0.50	1.23	0.13	0.99
	β_4	0.50	0.50	1.65	0.20	1.00
	β_1	1.50	-0.50	0.89	0.10	0.99
	β_2	1.50	0.50	1.19	0.13	1.00
	β_3	1.50	0.50	1.23	0.13	0.99
	β_4	1.50	0.50	1.65	0.20	1.00
	β_1	2.50	-0.50	0.88	0.09	0.99
	β_2	2.50	0.50	1.18	0.13	1.00
	β_3	2.50	0.50	1.22	0.13	1.00
	β_4	2.50	0.50	1.63	0.18	1.00

Table A.4: Coverage for ICP mixed effects models and true $\boldsymbol{\beta} = (-0.5, 0.5, 0.5, 0.5)'$

Model	Term	m	Truth	Avg.width	SD.width	Coverage
ICP	β_1	0.50	-0.50	1.41	0.31	0.23
	β_2	0.50	0.50	1.74	0.38	0.94
	β_3	0.50	0.50	1.85	0.43	0.95
	β_4	0.50	0.50	2.27	0.53	0.98
	β_1	1.50	-0.50	1.41	0.31	0.23
	β_2	1.50	0.50	1.74	0.38	0.94
	β_3	1.50	0.50	1.85	0.43	0.95
	β_4	1.50	0.50	2.27	0.53	0.98
	β_1	2.50	-0.50	1.59	0.38	0.21
	β_2	2.50	0.50	1.91	0.41	0.94
	β_3	2.50	0.50	2.04	0.50	0.95
	β_4	2.50	0.50	2.43	0.55	0.98
ICP ZIP	β_1	0.50	-0.50	1.36	0.25	0.97
	β_2	0.50	0.50	1.63	0.27	0.99
	β_3	0.50	0.50	1.67	0.29	0.96
	β_4	0.50	0.50	1.99	0.33	0.99
	β_1	1.50	-0.50	1.36	0.25	0.97
	β_2	1.50	0.50	1.63	0.27	0.99
	β_3	1.50	0.50	1.67	0.29	0.96
	β_4	1.50	0.50	1.99	0.33	0.99
	β_1	2.50	-0.50	1.52	0.34	0.96
	β_2	2.50	0.50	1.80	0.35	0.98
	β_3	2.50	0.50	1.83	0.36	0.96
	β_4	2.50	0.50	2.17	0.39	0.99

BIBLIOGRAPHY

Barouch, D. H., O'Brien, K. L., Simmons, N. L., King, S. L., Abbink, P., Maxfield, L. F., Sun, Y.-H., Porte, A. L., Riggs, A. M., Lynch, D. M., Clark, S. L., Backus, K., Perry, J. R., Seaman, M. S., Carville, A., Mansfield, K. G., Szinger, J. J., Fischer, W., and Korber, M. M. B. (2010), "Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys," *Nature*, 16, 319–323.

Ekanem, E., Dupont, H., Pickering, L., Selwyn, B., and Hawkins, C. (1983), "Transmission dynamics of enteric bacteria in the day-care centers," *American Journal of Epidemiology*, 118, 562–572.

Famoye, F. and Wang, W. (2004), "Censored generalized Poisson regression model," *Computational Statistics & Data Analysis*, 46, 547–560.

Fischer, W., Perkins, S., Theiler, J., Bhattacharya, T., Yusim, K., Funkhouser, R., Kuiken, C., Haynes, B., Letvin, N. L., Walker, B. D., Hahn, B. H., and Korber, B. T. (2007), "Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants," *Nature Medicine*, 13, 100–106.

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–533.

Kong, W.-P., Wu, L., Wallstrom, T. C., Fischer, W., Yang, Z.-Y., Ko, S.-Y., Letvin, N. L., Haynes, B. F., Hahn, B. H., Korber, B., , and Nabel, G. J. (2009), "Expanded Breadth of the T-Cell Response to Mosaic Human Immunodeficiency Virus Type 1 Envelope DNA Vaccination," *Journal of Virology*, 83, 2201–2215.

Long, J. and Freese, J. (2006), *Regression Models for Categorical and Limited Dependent Variables Using Stata. 2nd ed*, Stata Press.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall.

Miyahira, Y., Murata, K., Rodriguez, D. R. J. R., Esteban, M., Rodrigues, M. M., and Zavala, F. (1995), "Quantification of antigen specific CD8+ T cells using an ELISPOT assay," *Journal of Immunological Methods*, 181, 45–54.

Olsen, G. J., Matusda, H., Hagstrom, R., and Overbeek, R. (1994), "FastDnaML: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood." *Computer Applications in the Biosciences*, 10, 41–48.

Pruszynski, J. E. (2010), "Bayesian Models for Discrete Censored Sampling and Dose Finding," Ph.D. thesis, Baylor University.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rubin, D. (1987), *Multiple imputation for non response in surveys.*, Wiley.

Santra, S., Korber, B. T., Muldoon, M., Barouch, D. H., Nabel, G. J., Gao, F., Hahn, B. H., Haynes, B. F., and Letvin, N. L. (2008), "A centralized gene-based HIV-1 vaccine elicits broad cross-clade cellular immune responses in rhesus monkeys," *Proceedings of the National Academy of Sciences*, 105, 10489–10494.

Santra, S., Liao, H.-X., Zhang, R., Muldoon, M., Watson, S., Fischer, W., Theiler, J., Szinger, J., Balachandran, H., Buzby, A., Quinn, D., Parks, R. J., Tsao, C.-Y., Carville, A., Mansfield, K. G., Pavlakis, G. N., Felber, B. K., Haynes, B. F., and Letvin, B. T. K. N. L. (2010), "Mosaic vaccines elicit CD8+ T lymphocyte responses that confer enhanced immune coverage of diverse HIV strains in monkeys," *Nature Medicine*, 16, 324–329.

Simek, M. D., Rida, W., Priddy, F. H., Pung, P., Carrow, E., Laufer, D. S., Lehrman, J. K., Boaz, M., Tarragona-Fiol, T., Miiro, G., Birungi, J., Pozniak, A., McPhee, D. A., Manigart, O., Karita, E., Inwoley, A., Jaoko, W., Dehovitz, J., Bekker, L.-G., Pittsuttithum, P., Paris, R., Walker, L. M., Poignard, P., Wrin, T., Fast, P. E., Burton, D. R., and Koff, W. C. (2009), "Human Immunodeficiency Virus Type 1 Elite Neutralizers: Individuals with Broad and Potent Neutralizing Activity Identified by Using a High-Throughput Neutralization Assay together with an Analytical Selection Algorithm," *Journal of Virology*, 83, 7337–7348.

Terza, J. V. (1985), "A Tobit-type estimator for the censored Poisson regression model," *Economics Letters*, 18, 361–365.