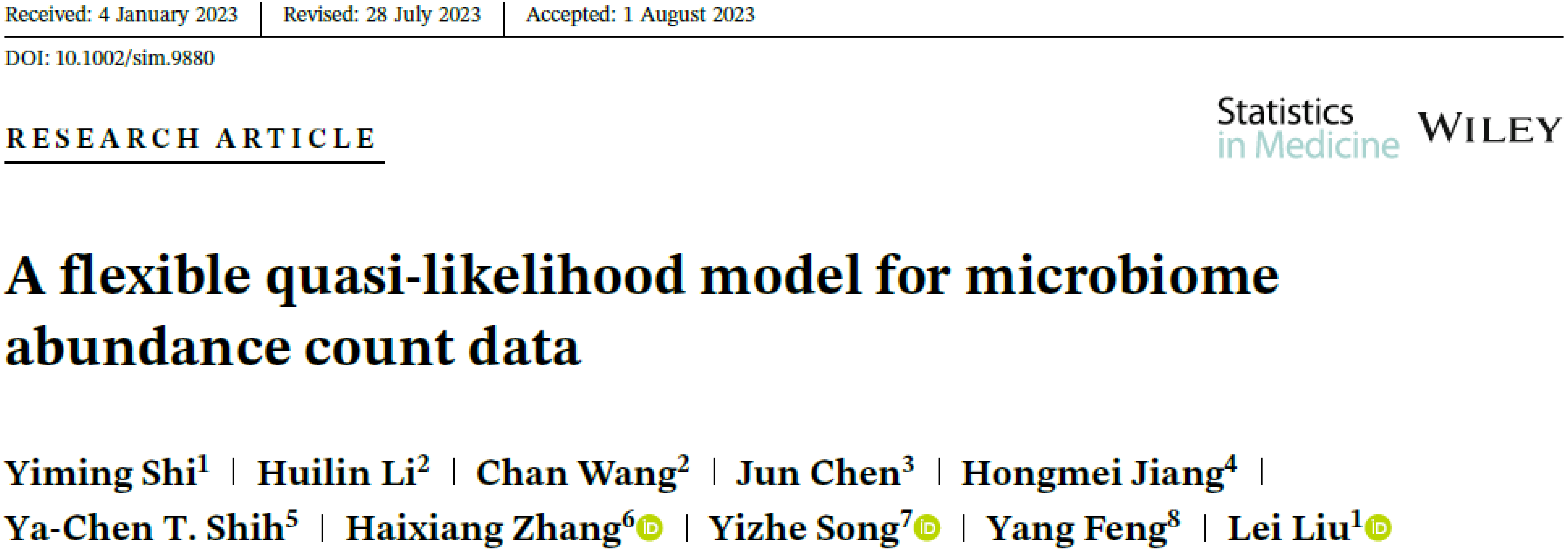
BIBS Seminar



Received: 4 January 2023 | Revised: 28 July 2023 | Accepted: 1 August 2023

DOI: 10.1002/sim.9880

RESEARCH ARTICLE

Statistics in Medicine WILEY

A flexible quasi-likelihood model for microbiome abundance count data

Yiming Shi[1] | Huilin Li[2] | Chan Wang[2] | Jun Chen[3] | Hongmei Jiang[4] | Ya-Chen T. Shih[5] | Haixiang Zhang[6] | Yizhe Song[7] | Yang Feng[8] | Lei Liu[1]

Presenter: Mozaffar Hosain

21 December 2023

Contents

- Introduction
- Basic concepts of quasi-likelihood
- Methods
- Simulation
- Application
- Discussion

Introduction

- The collection of complete set of genomes from all the microbes is referred to as microbiome.
- Several diseases, including obesity, diabetes, Crohn’s disease, bacterial vaginosis, and cancer, among others are associated with microbiome profile.
- Recently, microbiome studies have been growing rapidly due to next-generation sequencing (NGS) technologies.
- The abundance count for a microbial taxon is often sparse and overdispersed.
- Methods based on the normality assumption are typically inadequate and often result in invalid inferences.

Introduction

- A quasi-likelihood approach is used in this paper to address the overdispersion usually observed in microbiome data.

- Quasi-likelihood can be used as an alternative to maximum likelihood estimation for generalized liner models (GLM).

- In this framework, the response variable does not assume any distributional form.

- Only the first two moments (mean and variance), and the relationship between them are needed.

Introduction

- A nonparametric function of the nonlinear associations between mean and variance structure modeled by penalized splines was considered by Chen et al.(2013).

- Chiou and Muller (1999) proposed a nonparametric quasi-likelihood method, with a nonparametric link function and a nonparametric variance-mean relationship.

- In this article, a flexible quasi-likelihood (FQL) approach is adapted for microbiome data, motivated by the aforementioned two methods.

- Performance of FQL is compared with other available methods.

- An R package "**fql**" is also developed to implement the proposed method.

Quasi-likelihood

- To construct a likelihood function, it is necessary to know the probability distributions of the random variables.
- In some situations, the underlying probability distribution is not known.
- Also, in other situations, the assumed distribution may be inadequate.
- Another possibility is that the underlying theoretical model may be too complicated to permit parameter estimation and statistical inference.

Quasi-likelihood

- However, we may still have substantial information about the data, such as:
  - type of response (discrete, continuous, nonnegative, symmetric, skewed, etc.)
  - whether or not the observations are statistically independent
  - mean and variance relationship
  - the possible nature of the relationship between the mean response and one or more covariates
- In such cases, quasi-likelihood is a method for statistical inference when it is not possible to construct a likelihood function.

Quasi-likelihood

- Let $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ be a vector of independent random variables with mean vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$.

- Let $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)'$ be a vector of unknown parameters $p \leq n$.

- We assume that the parameters of interest, $\boldsymbol{\beta}$, relate to the dependence of $\boldsymbol{\mu}$ on covariates $\boldsymbol{x}$.

- It is denoted by the notation that Y_i has a mean of $\mu_i(\beta)$.

- Also, we assume that $Var(Y_i) = \phi V(\mu_i)$, where $V(.)$ is a known function and ϕ is a possibly unknown scale parameter.

- Hence, $Var(\boldsymbol{Y}) = \phi V(\boldsymbol{\mu})$, where $V(\boldsymbol{\mu}) = diag\{V(\mu_1), \dots, V(\mu_n)\}$.

Quasi-likelihood

Construction of quasi-likelihood function:

- Let us define the random variable $U_i = \frac{Y_i - \mu_i}{\phi V(\mu_i)}$.
- U_i has the following properties like a score function:

$$E(U_i) = 0, Var(U_i) = E\left(U_i^2\right) = \frac{E\left[(Y_i - \mu_i)^2\right]}{[\phi V(\mu_i)]^2} = \frac{1}{\phi V(\mu_i)}, \text{ and}$$

$$E\left(\frac{\partial U_i}{\partial \mu_i}\right) = -Var(U_i).$$

- The quasi-likelihood for μ_i based on data y_i is defined as:

$$Q(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt.$$

- So, the quasi-likelihood for the independent observations $Y_1, \dots, Y_n$ is:

$$\mathrm{Q}(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} Q(\mu_i; y_i).$$

Quasi-likelihood

Quasi-likelihood estimating equations:

- To estimate β_j,

$$0 = \frac{\partial \mathrm{Q}(\boldsymbol{\mu}; \boldsymbol{y})}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{\partial Q(\mu_i; y_i)}{\partial \beta_j}$$

$$= \sum_{i=1}^{n} \frac{\partial Q(\mu_i; y_i)}{\partial \mu_i} \left(\frac{\partial \mu_i}{\partial \beta_j} \right)$$

$$= \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\phi V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \beta_j} \right)$$

- In matrix notation, $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} = \boldsymbol{D}_{n \times p}$, where the (i, j) component of $\boldsymbol{D}$ is $\frac{\partial \mu_i}{\partial \beta_j}$.

- The estimating equation is then $U(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$,

where $U(\boldsymbol{\beta}) = \boldsymbol{D}' V^{-\mathbf{1}} \frac{(\boldsymbol{y} - \boldsymbol{\mu})}{\phi}$, is called the quasi-score function.

Methods

Model

- Let Y_i be the count in sample i $(i = 1, 2, \ldots, n)$ for a taxon.
- Let $E(Y_i) = \mu_i$, $\boldsymbol{y} = (Y_1, \ldots, Y_n)$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$.
- The following model is defined:

$$\log(\mu_i) = \boldsymbol{X}_i'\boldsymbol{\beta} \qquad (1)$$

$$Var(Y_i) = V(\mu_i), \qquad (2)$$

where $\boldsymbol{X}_i$ is a vector of p covariates, $\boldsymbol{\beta}$ is the corresponding vector of their regression coefficients, and $V(\cdot)$ is a function of μ_i.

Methods

Model

- A log link function is considered in model (1).
- The variance is modeled as an unknown function $V(\cdot)$ of the mean in (2).
- As a particular case of (2), $V(\mu_i) = \mu_i + \frac{\mu_i^2}{r}$ (r is the dispersion parameter) can be considered for the negative binomial distribution.
- In this article, no additional assumptions about the distribution of the response variable were made to bring up the model more flexible and widely applicable.

Methods

Estimation and Inference

- The flexible quasi-likelihood (FQL) function is defined as:

$$Q(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt \qquad (3)$$

- Quasi score equation to estimate the parameter vector $\boldsymbol{\beta}$:

$$U^* = \sum_{i=1}^{n} \boldsymbol{D}_i V^{-1}(\mu_i)(y_i - \mu_i) = \mathbf{0}, \qquad (4)$$

where $\mu_i = e^{\boldsymbol{X}_i' \boldsymbol{\beta}}$, and $\boldsymbol{D}_i = \frac{d\mu_i}{d\beta} = \mu_i \boldsymbol{X}_i$ is a $p \times 1$ vector.

Methods

Estimation and Inference

- Estimation procedure:

1. Initialize $\boldsymbol{\beta}$ by fitting a model assuming a constant $V(\mu_{\mathrm{i}})$ for all subjects. Set $\mu_i = e^{\boldsymbol{X}_i'\boldsymbol{\beta}}$.

2. Estimate the unknown variance function $V(\mu_i)$ by minimizing the penalized least squares function:

$$\sum_i[(y_i - \hat{\mu}_i)^2 - V(\hat{\mu}_i)]^2 + J_\lambda\big(V(\hat{\mu}_i)\big), \qquad (5)$$

➢ where $J_\lambda\big(V(\hat{\mu}_i)\big)$ is a penalty function with parameter λ

➢ P-spline with quadratic penalty is used to estimate $V(\mu_i)$

Methods

Estimation and Inference

- Estimation procedure:
  - ➢ $J_\lambda\left(V(\hat{\mu}_i)\right) = \lambda \sum_i \boldsymbol{\alpha}' \boldsymbol{S}_i \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the vector of parameters in the P-spline model of $V(\hat{\mu}_i)$.
  - ➢ $\boldsymbol{S}_i$ is a positive semi-definite matrix.

3. Estimate $\boldsymbol{\beta}$ by solving the quasi score Equation (4). The Newton-Raphson method with Fisher scoring gives the following estimate of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + \left[\sum_{i=1}^n \widehat{\boldsymbol{D}}_i^{(k)} \left(\widehat{\boldsymbol{D}}_i^{(k)}\right)' \left(\widehat{V}_i^{(k)}\right)^{-\mathbf{1}}\right]^{-1} \left[\sum_{i=1}^n \widehat{\boldsymbol{D}}_{\boldsymbol{i}}^{(k)} \left(y_i - \hat{\mu}_i^{(k)}\right) \left(\widehat{V}_i^{(k)}\right)^{-\mathbf{1}}\right], \text{ and}$$

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}^{(k+1)}\right) = \left[\sum_{i=1}^n \widehat{\boldsymbol{D}}_i \widehat{\boldsymbol{D}}_i' \widehat{V}_i^{-1}\right]^{-1} \left[\sum_{i=1}^n \widehat{\boldsymbol{D}}_i \widehat{\boldsymbol{D}}_i' (y_i - \widehat{\mu}_i)^2 \widehat{V}_i^{-2}\right] \left[\sum_{i=1}^n \widehat{\boldsymbol{D}}_i \widehat{\boldsymbol{D}}_i' \widehat{V}_i^{-1}\right]^{-1}.$$

Simulation

1. Data were generated from the following distributions:

- ➢ Negative binomial
- ➢ Poisson
- ➢ Gamma
- ➢ Pareto

- The Gamma and Pareto are continuous distributions, these are considered as mis-specified distributions, and the rounded values are taken as the count outcome.
- For each of the above distributions, 600 datasets were generated with sample size $n = 400$.

Simulation

- A single covariate X is included ($X \sim U(0,1)$.
- For all of the 4 distributions, the following model was considered:

$$\log(\mu) = \beta_0 + \beta_1 x$$

- Parameter settings for the type I error rate: $\beta_0 = 1$, and $\beta_1 = 0$.
- Parameter settings for power: $\beta_0 = 1$, and $\beta_1 = 0.2$.

Simulation

Example 1: Negative binomial distribution

- The probability function:

$$f(y;r,p) = \frac{\Gamma(y+r)}{\Gamma(r)y!} p^r (1-p)^y \text{ for } y = 0, 1, 2, . \ . \ .,$$

where r is the number of successes, k is the number of failures, and p is the probability of success on each trial.

- Mean $(\mu) = \frac{r(1-p)}{p}$, Variance $(\sigma^2) = \frac{r(1-p)}{p^2}$, hence $\sigma^2 = \mu + \frac{\mu^2}{r}$.
- The value of r is fixed at $r = 2$.
- Then Y values are generated with mean μ such that $\log(\mu) = \beta_0 + \beta_1 x$, with $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_1 = 0.2$.

Simulation

Example 2: Poisson distribution

- Probability function:

$$f(y;\mu) = \frac{e^{-\mu}}{\mu^{y} y!} \text{ for } y = 0, 1, 2, . \ . \ .,$$

where mean = variance = μ.

- Y values are generated with mean μ, where $\log(\mu) = \beta_0 + \beta_1 x$, with $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_1 = 0.2$.

Simulation

Example 3: Mis-specified (gamma) distribution

- Probability function:

$$f(y; a, s) = \frac{e^{-sy} y^{a-1}}{s^a \Gamma(a)}, y > 0,$$

where a and s are the shape, and scale parameters respectively.

- $E\,(Y) = as, Var(Y) = as^2$.
- Parameter settings: $a = \mu^{3/2}$, and $s = \mu^{-1/2}$.
- Therefore, $E\,(Y) = \mu$, and $Var(Y) = \mu^{1/2}$.
- Then Y values are generated with μ, where $\log(\mu) = \beta_0 + \beta_1 x$, with $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_1 = 0.2$.

Simulation

Example 4: Mis-specified (Pareto) distribution

- Probability function:

$$f(y;\alpha,\beta) = \frac{\alpha\beta^{\alpha}}{y^{\alpha+1}}\,,\ \ y > \beta,$$

where α and β are the shape, and scale parameters respectively.

- $E\,(Y) = \frac{\alpha\beta}{\alpha-1}$, for $\alpha > 1$, and $Var(Y) = \frac{\alpha\beta^2}{(\alpha-2)(\alpha-1)^2}$ for $\alpha > 2$.

- Parameter settings: $\alpha = 2.5$, and $\beta = \frac{3}{5}\mu$.

- Therefore, $E\,(Y) = \mu$, and $Var(Y) = \frac{4}{5}\mu^2$.

- Then Y values are generated with μ, where $\log(\mu) = \beta_0 + \beta_1 x$, with $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_1 = 0.2$.

Simulation

- The following models are fitted to all datasets from the 4 distributions:
  - ➢ Flexible quasi-likelihood (FQL) model
  - ➢ Negative binomial GLM
  - ➢ Poisson GLM
- All 3 models give asymptotically unbiased results, and mean squared errors (MSE) from them are very close.

Simulation

Simulation Results (data simulated from NB distribution)

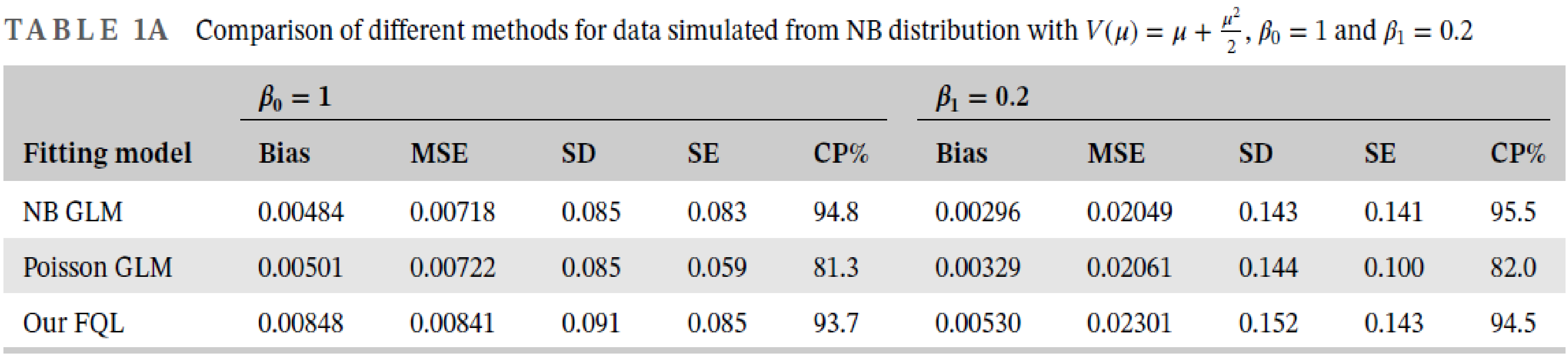


TABLE 1A Comparison of different methods for data simulated from NB distribution with $V(\mu) = \mu + \frac{\mu^2}{2}$, $\beta_0 = 1$ and $\beta_1 = 0.2$

	$\beta_0 = 1$					$\beta_1 = 0.2$				
Fitting model	**Bias**	**MSE**	**SD**	**SE**	**CP%**	**Bias**	**MSE**	**SD**	**SE**	**CP%**
NB GLM	0.00484	0.00718	0.085	0.083	94.8	0.00296	0.02049	0.143	0.141	95.5
Poisson GLM	0.00501	0.00722	0.085	0.059	81.3	0.00329	0.02061	0.144	0.100	82.0
Our FQL	0.00848	0.00841	0.091	0.085	93.7	0.00530	0.02301	0.152	0.143	94.5

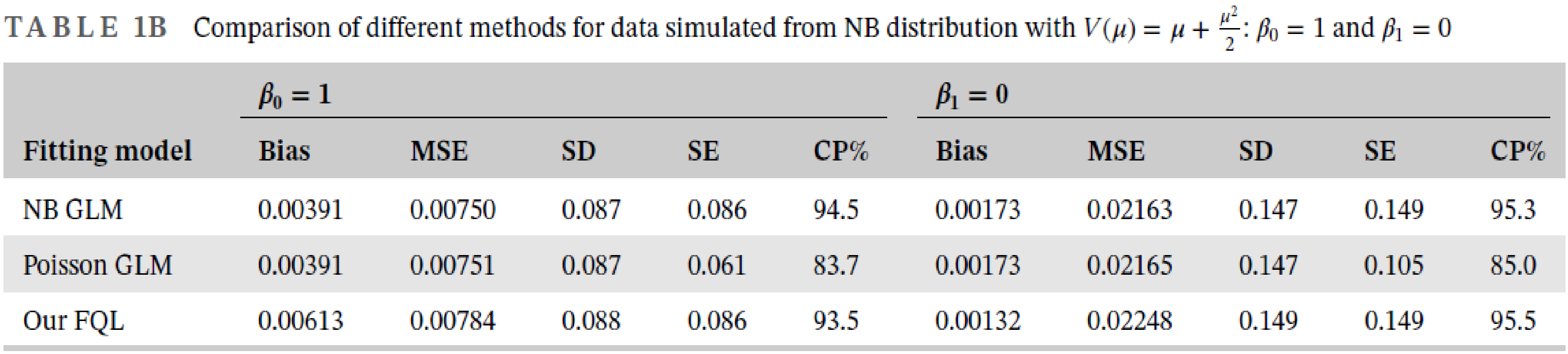


TABLE 1B Comparison of different methods for data simulated from NB distribution with $V(\mu) = \mu + \frac{\mu^2}{2}$: $\beta_0 = 1$ and $\beta_1 = 0$

	$\beta_0 = 1$					$\beta_1 = 0$				
Fitting model	**Bias**	**MSE**	**SD**	**SE**	**CP%**	**Bias**	**MSE**	**SD**	**SE**	**CP%**
NB GLM	0.00391	0.00750	0.087	0.086	94.5	0.00173	0.02163	0.147	0.149	95.3
Poisson GLM	0.00391	0.00751	0.087	0.061	83.7	0.00173	0.02165	0.147	0.105	85.0
Our FQL	0.00613	0.00784	0.088	0.086	93.5	0.00132	0.02248	0.149	0.149	95.5

- FQL and NB GLM both show the coverage probabilities close to the nominal level (95%)
- Under-coverage for Poisson GLM.

Simulation

Simulation Results (data simulated from Poisson distribution)

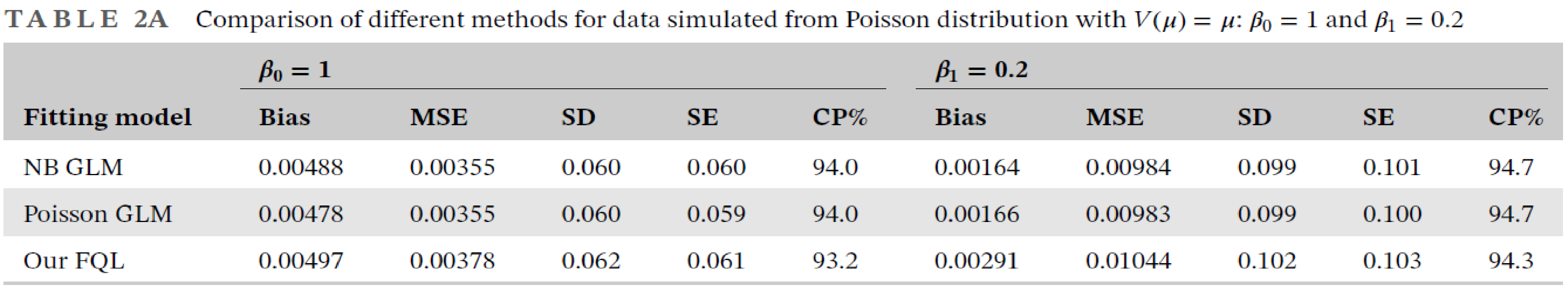


TABLE 2A Comparison of different methods for data simulated from Poisson distribution with $V(\mu) = \mu$: $\beta_0 = 1$ and $\beta_1 = 0.2$

	$\beta_0 = 1$					$\beta_1 = 0.2$				
Fitting model	**Bias**	**MSE**	**SD**	**SE**	**CP%**	**Bias**	**MSE**	**SD**	**SE**	**CP%**
NB GLM	0.00488	0.00355	0.060	0.060	94.0	0.00164	0.00984	0.099	0.101	94.7
Poisson GLM	0.00478	0.00355	0.060	0.059	94.0	0.00166	0.00983	0.099	0.100	94.7
Our FQL	0.00497	0.00378	0.062	0.061	93.2	0.00291	0.01044	0.102	0.103	94.3

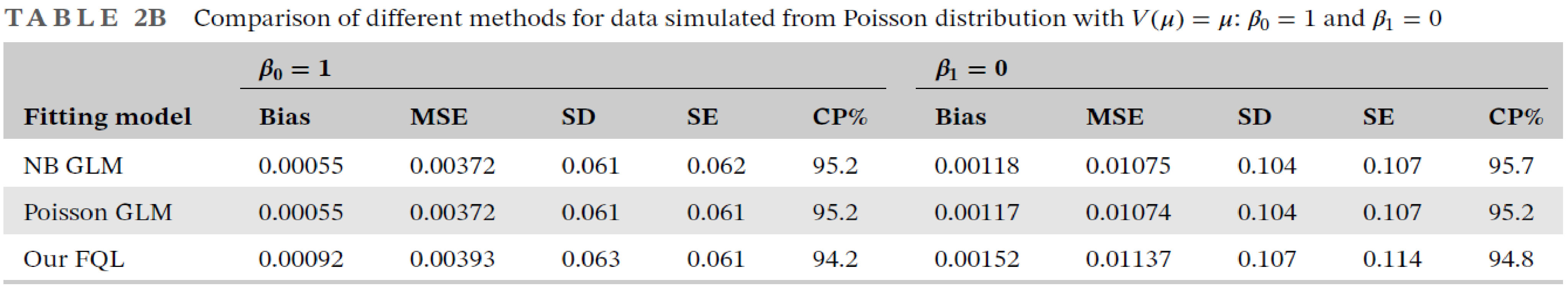


TABLE 2B Comparison of different methods for data simulated from Poisson distribution with $V(\mu) = \mu$: $\beta_0 = 1$ and $\beta_1 = 0$

	$\beta_0 = 1$					$\beta_1 = 0$				
Fitting model	**Bias**	**MSE**	**SD**	**SE**	**CP%**	**Bias**	**MSE**	**SD**	**SE**	**CP%**
NB GLM	0.00055	0.00372	0.061	0.062	95.2	0.00118	0.01075	0.104	0.107	95.7
Poisson GLM	0.00055	0.00372	0.061	0.061	95.2	0.00117	0.01074	0.104	0.107	95.2
Our FQL	0.00092	0.00393	0.063	0.061	94.2	0.00152	0.01137	0.107	0.114	94.8

- All 3 models perform well in terms of coverage probabilities.

Simulation

Simulation Results (data simulated from mis-specified Gamma distribution)

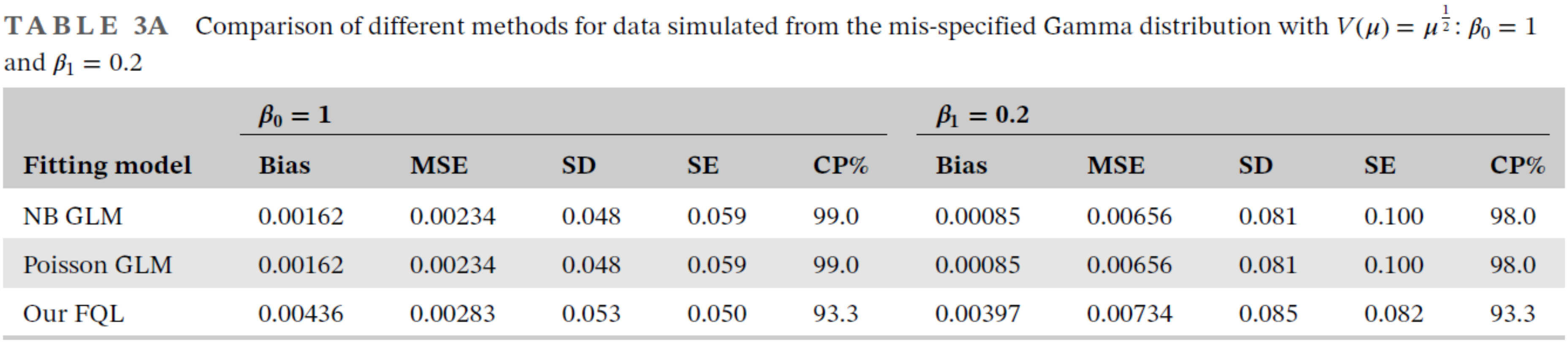


TABLE 3A Comparison of different methods for data simulated from the mis-specified Gamma distribution with $V(\mu) = \mu^{\frac{1}{2}}$: $\beta_0 = 1$ and $\beta_1 = 0.2$

	$\beta_0 = 1$					$\beta_1 = 0.2$				
Fitting model	**Bias**	**MSE**	**SD**	**SE**	**CP%**	**Bias**	**MSE**	**SD**	**SE**	**CP%**
NB GLM	0.00162	0.00234	0.048	0.059	99.0	0.00085	0.00656	0.081	0.100	98.0
Poisson GLM	0.00162	0.00234	0.048	0.059	99.0	0.00085	0.00656	0.081	0.100	98.0
Our FQL	0.00436	0.00283	0.053	0.050	93.3	0.00397	0.00734	0.085	0.082	93.3

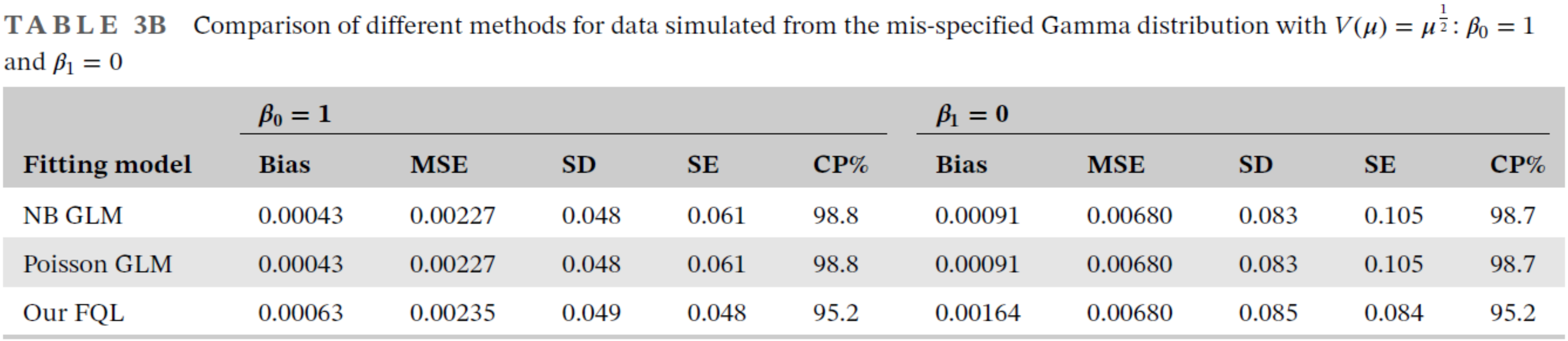


TABLE 3B Comparison of different methods for data simulated from the mis-specified Gamma distribution with $V(\mu) = \mu^{\frac{1}{2}}$: $\beta_0 = 1$ and $\beta_1 = 0$

	$\beta_0 = 1$					$\beta_1 = 0$				
Fitting model	**Bias**	**MSE**	**SD**	**SE**	**CP%**	**Bias**	**MSE**	**SD**	**SE**	**CP%**
NB GLM	0.00043	0.00227	0.048	0.061	98.8	0.00091	0.00680	0.083	0.105	98.7
Poisson GLM	0.00043	0.00227	0.048	0.061	98.8	0.00091	0.00680	0.083	0.105	98.7
Our FQL	0.00063	0.00235	0.049	0.048	95.2	0.00164	0.00680	0.085	0.084	95.2

- Both the NB and Poisson GLMs overestimate the standard errors, resulting in over-coverage.
- The FQL performs the best with reasonable coverage probabilities.

Simulation

Simulation Results (data simulated from mis-specified Pareto distribution)

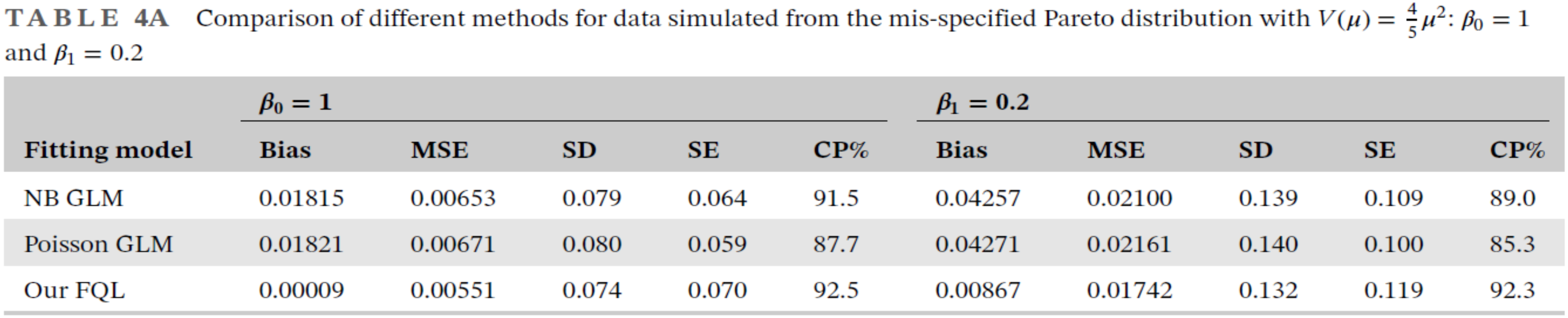


TABLE 4A Comparison of different methods for data simulated from the mis-specified Pareto distribution with $V(\mu) = \frac{4}{5}\mu^2$: $\beta_0 = 1$ and $\beta_1 = 0.2$

	$\beta_0 = 1$					$\beta_1 = 0.2$				
Fitting model	**Bias**	**MSE**	**SD**	**SE**	**CP%**	**Bias**	**MSE**	**SD**	**SE**	**CP%**
NB GLM	0.01815	0.00653	0.079	0.064	91.5	0.04257	0.02100	0.139	0.109	89.0
Poisson GLM	0.01821	0.00671	0.080	0.059	87.7	0.04271	0.02161	0.140	0.100	85.3
Our FQL	0.00009	0.00551	0.074	0.070	92.5	0.00867	0.01742	0.132	0.119	92.3

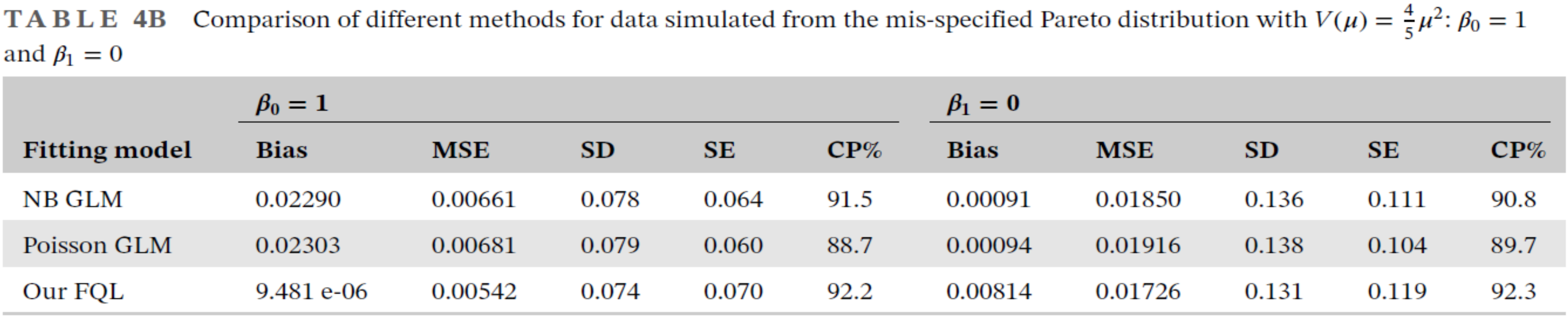
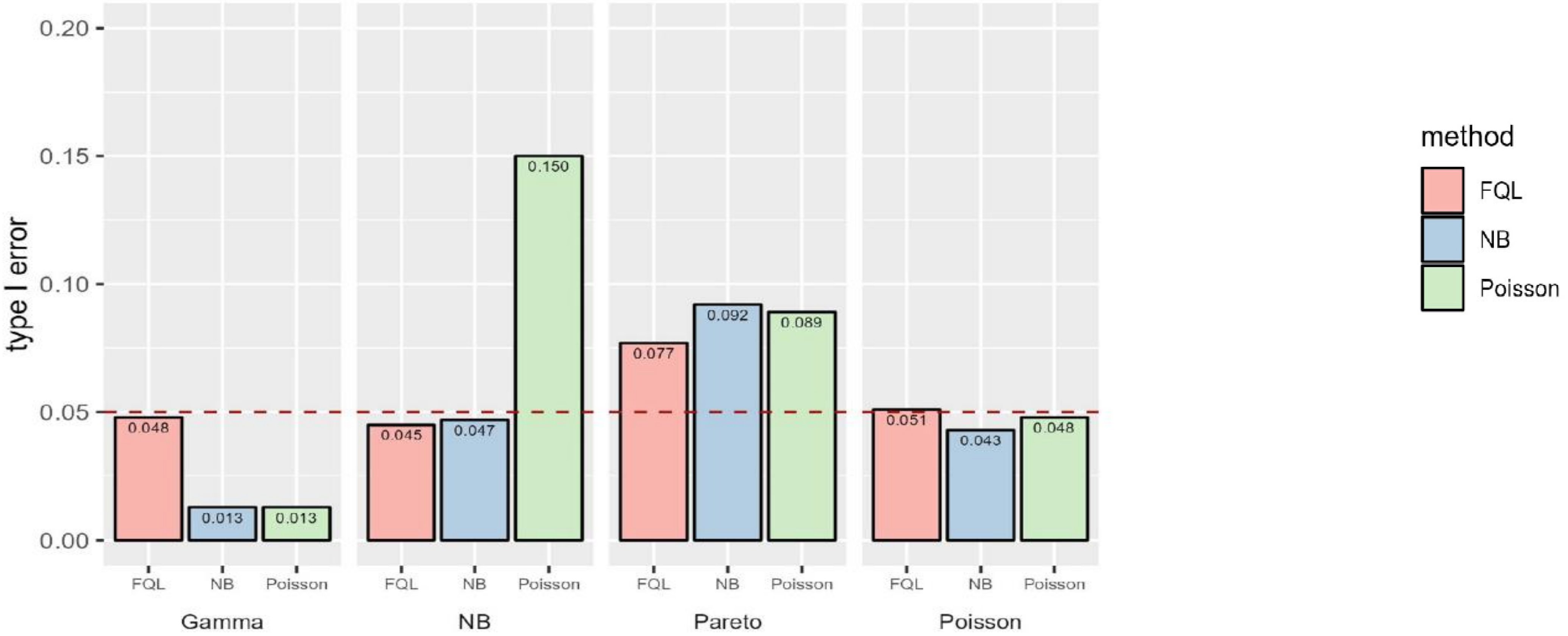


TABLE 4B Comparison of different methods for data simulated from the mis-specified Pareto distribution with $V(\mu) = \frac{4}{5}\mu^2$: $\beta_0 = 1$ and $\beta_1 = 0$

	$\beta_0 = 1$					$\beta_1 = 0$				
Fitting model	**Bias**	**MSE**	**SD**	**SE**	**CP%**	**Bias**	**MSE**	**SD**	**SE**	**CP%**
NB GLM	0.02290	0.00661	0.078	0.064	91.5	0.00091	0.01850	0.136	0.111	90.8
Poisson GLM	0.02303	0.00681	0.079	0.060	88.7	0.00094	0.01916	0.138	0.104	89.7
Our FQL	9.481 e-06	0.00542	0.074	0.070	92.2	0.00814	0.01726	0.131	0.119	92.3

- Both the NB and Poisson GLMs underestimate the standard errors, resulting smaller coverage probabilities than that of the proposed FQL model.

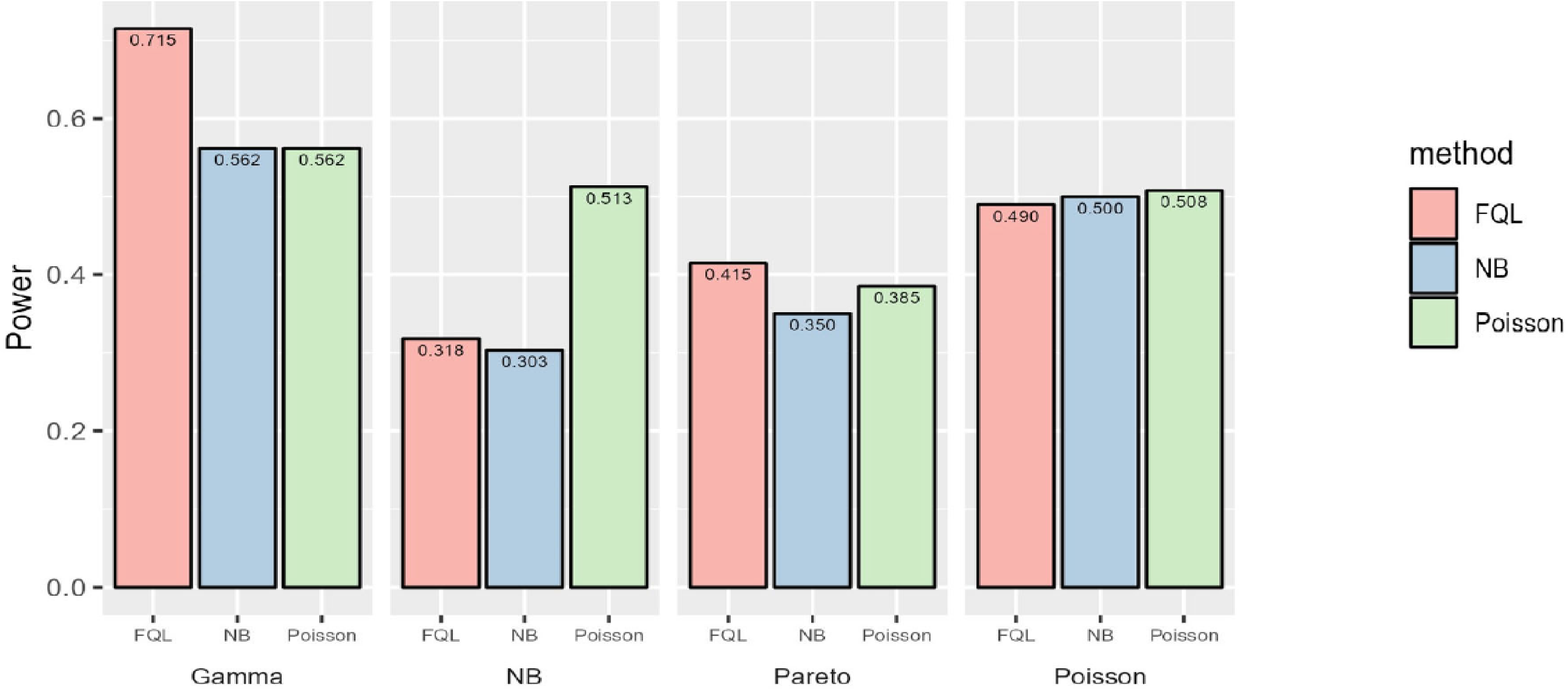
Simulation Results (Type I error)



- When data are generated from the NB distribution, Poisson GLM cannot control the type I error rate.
- For the data from gamma distribution, FQL has better type I error control, NB and Poisson GLMs are more conservative.
- For Pareto, the proposed FQL gives the lowest type I error rate.

Simulation

Simulation Results (Power)



- When data are generated from the gamma, and Pareto distributions, the proposed FQL model shows the highest power than the other two models.
- That is, when the underlying distribution is mis-specified, the NB and Poisson GLMs models may produce misleading results.

Simulation

2. Simulation from real data

- A real data based simulation framework for data generation is used.
- The simulation framework captures the complexity of microbiome data by generating random samples from a large reference dataset and using these reference samples as templates to generate new samples.
- A real dataset is used as the reference data.
- The performance of the proposed model FQL was compared with NB, Poisson GLMs, and with ZicoSeq (zero-inflated compositional sequencing) model.

Simulation

2. Simulation from real data

- 400 samples were generated for each simulation with 100 operational taxonomic units (OTUs).
- 20 of the OTUs are differentially expressed.
- The abundance for the differentially expressed OUT is:

$$C_i' = C_i exp(\beta_1 X_i + \varepsilon_i),$$

where $X_i \sim U(0,1)$, $\beta_1 = 0.2$, and C_i is the random abundance from the reference real data.

- Based on the abundance, the OTUs are grouped as

Top half of the abundance range: Common OTUs

bottom half of the abundance range: Rare OTUs

Simulation

2. Simulation from real data

- For the preprocessing of the simulation reference dataset, OTUs with prevalence less than 25% are excluded.
- The classification of OTUs is then:

Prevalence from 100% to 62.5%: Common

prevalence ranging from 62.5% to 25%: Rare

Simulation

Simulation results (simulation from real data)

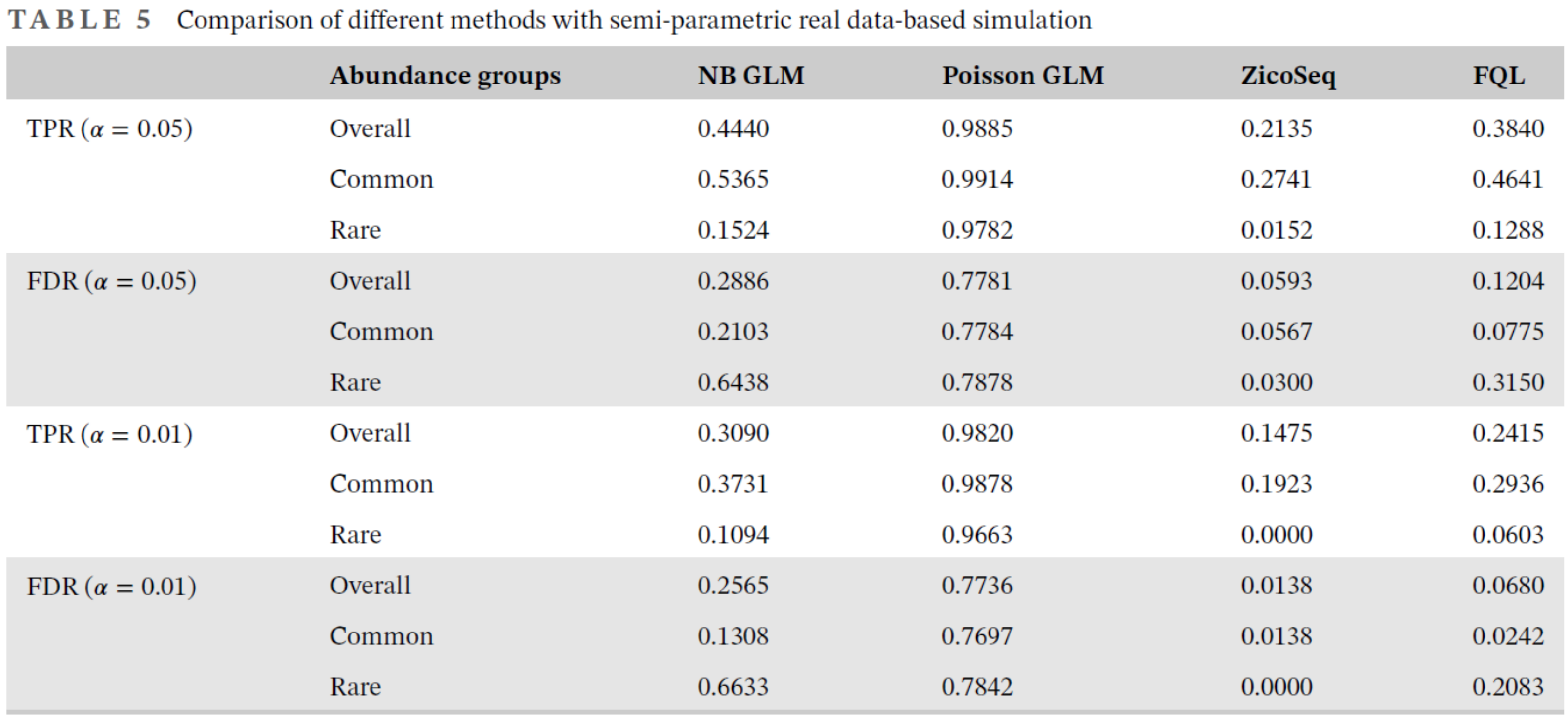


TABLE 5 Comparison of different methods with semi-parametric real data-based simulation

	Abundance groups	NB GLM	Poisson GLM	ZicoSeq	FQL
TPR ($\alpha = 0.05$)	Overall	0.4440	0.9885	0.2135	0.3840
	Common	0.5365	0.9914	0.2741	0.4641
	Rare	0.1524	0.9782	0.0152	0.1288
FDR ($\alpha = 0.05$)	Overall	0.2886	0.7781	0.0593	0.1204
	Common	0.2103	0.7784	0.0567	0.0775
	Rare	0.6438	0.7878	0.0300	0.3150
TPR ($\alpha = 0.01$)	Overall	0.3090	0.9820	0.1475	0.2415
	Common	0.3731	0.9878	0.1923	0.2936
	Rare	0.1094	0.9663	0.0000	0.0603
FDR ($\alpha = 0.01$)	Overall	0.2565	0.7736	0.0138	0.0680
	Common	0.1308	0.7697	0.0138	0.0242
	Rare	0.6633	0.7842	0.0000	0.2083

- FQL gives the FDR values which are comparable to that of ZicoSeq (smaller than the NB and Poisson GLMs).
- FQL has a TPR close to NB GLMs which is much higher than that of ZicoSeq.

Application

Real data analysis

- A study based on real data was conducted for the early events of carcinogenesis by investigating shifts in the gut microbiota of patients with adenomas.
- The data contained fecal microbiota information of 800 patients.
- Patients with adenomas ($n = 266$) and without ($n = 534$).
- Total number of OTUs (genus level): 178.
- To consider different zero-inflation status, the taxa with prevalence less than 15% (76 OTUs left), and 25% (63 OTUs left) were excluded.
- The objective is to study the effect of adenomas on the abundance of these OTUs.

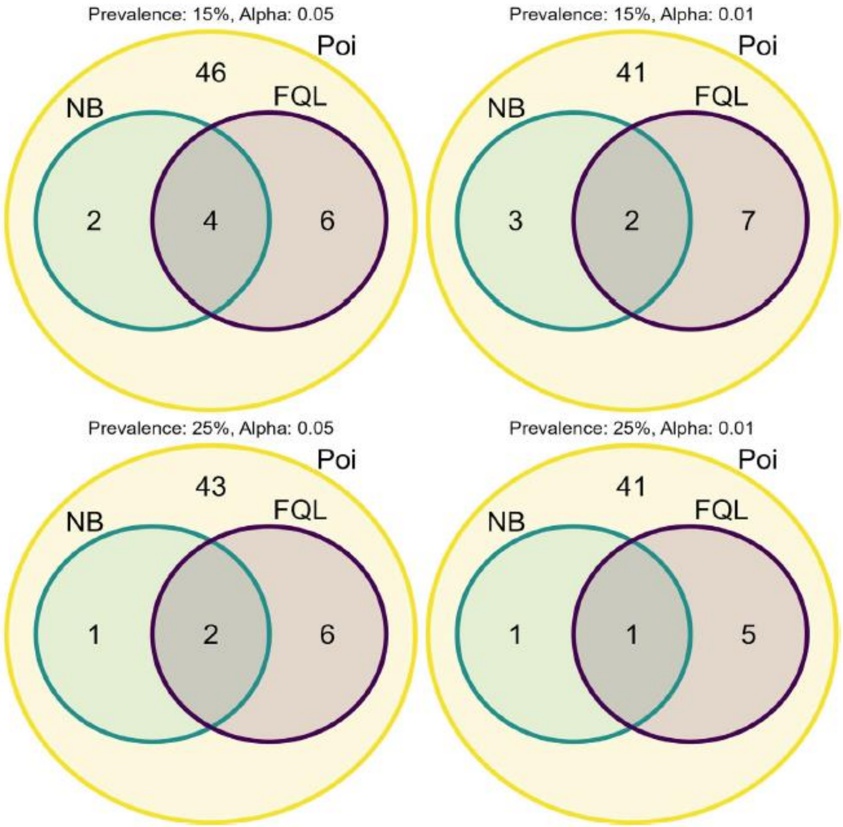
Application

Real data analysis

- Four models were applied to these real data:
    - ➢ The proposed FQL model
    - ➢ NB GLM
    - ➢ Poisson GLM
    - ➢ ZicoSeq model
- Covariates used in the models: gender, ever smoking, having polyps or not, and sequencing batch

Application

Real data analysis: Results



Prevalence: 15%, Alpha: 0.05
Poi
46
NB
FQL
2
4
6
Prevalence: 15%, Alpha: 0.01
Poi
41
NB
FQL
3
2
7
Prevalence: 25%, Alpha: 0.05
Poi
43
NB
FQL
1
2
6
Prevalence: 25%, Alpha: 0.01
Poi
41
NB
FQL
1
1
5

FIGURE 2 The Venn diagrams for different prevalence cutoffs (15% vs 25%) and different significance (0.05 vs 0.01)

- ZicoSeq did not identify any differentially abundance taxa.
- Poisson GLM gives the largest number of significant OTUs under all scenarios (consistent with simulation results: inflated type I error rate).
- FQL identified more OTUs than the NB GLM. Which justified the simulation results (more powerful and identified more OTUs when the underlying distribution is mis-specified).
- The results are robust under different zero-inflation levels and significance levels.

Application

Real data analysis: Results

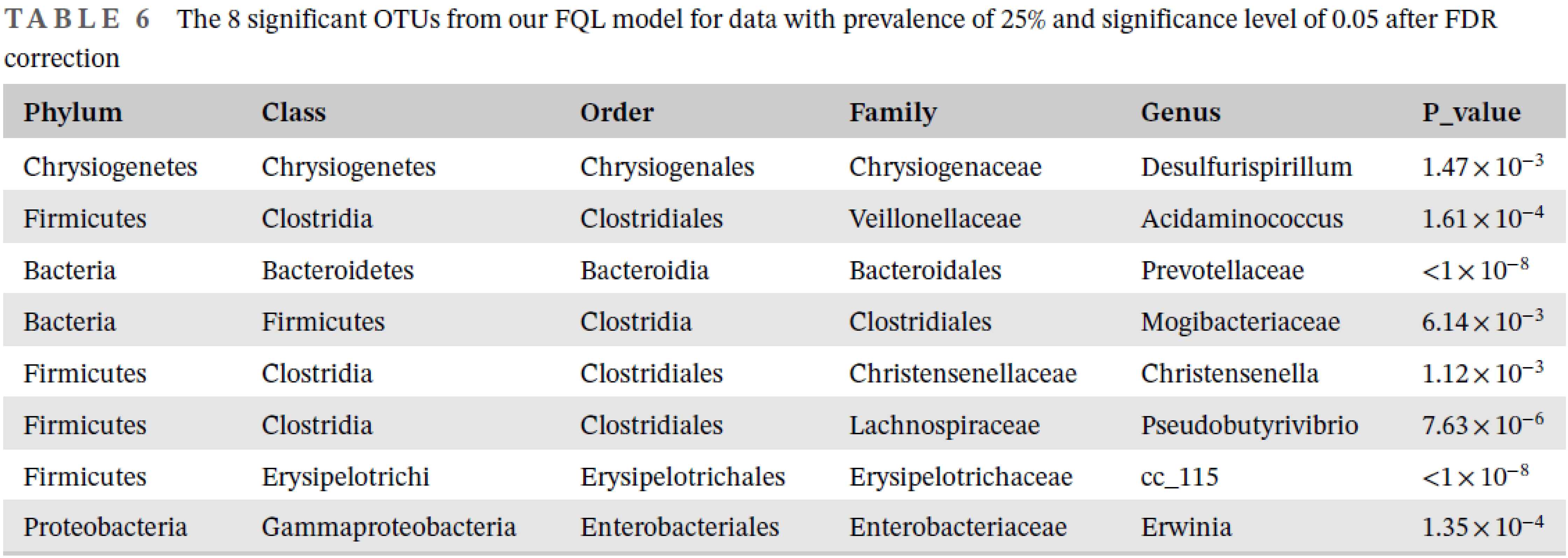


TABLE 6 The 8 significant OTUs from our FQL model for data with prevalence of 25% and significance level of 0.05 after FDR correction

Phylum	Class	Order	Family	Genus	P_value
Chrysiogenetes	Chrysiogenetes	Chrysiogenales	Chrysiogenaceae	Desulfurispirillum	1.47×10^{-3}
Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Acidaminococcus	1.61×10^{-4}
Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	$<1 \times 10^{-8}$
Bacteria	Firmicutes	Clostridia	Clostridiales	Mogibacteriaceae	6.14×10^{-3}
Firmicutes	Clostridia	Clostridiales	Christensenellaceae	Christensenella	1.12×10^{-3}
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Pseudobutyrivibrio	7.63×10^{-6}
Firmicutes	Erysipelotrichi	Erysipelotrichales	Erysipelotrichaceae	cc_115	$<1 \times 10^{-8}$
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Erwinia	1.35×10^{-4}

Discussion

- The FQL model does not need the specification of the distribution function, hence it is more robust to model mis-specification.
- Simulation, and real studies show that FQL has better performance than the competing models.
- The proposed model does not specifically address zero inflation, which leads to less satisfactory performance for rare taxa in the simulation study.
- If the percentage of zeros is very high, then a one-part model should be avoided.
- The model can be extended:
    - using other link functions (e.g., logit)
    - adding random effects to the model for clustered/longitudinal data
- To increase the efficiency, the phylogenetic or taxonomic tree structure among different taxa can be incorporated.

Thank you!