Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

doi:10.3934/xx.xx.xx.xx

pp. **X–XX**

A NUMERICAL METHOD TO COMPUTE FISHER INFORMATION FOR A SPECIAL CASE OF HETEROGENEOUS NEGATIVE BINOMIAL REGRESSION

Xin Guo

Department of Applied Mathematics,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong, China

Qiang Fu*

Department of Sociology,
The University of British Columbia,
V6T 1Z1, Vancouver, BC, Canada

Yue Wang

Department of Applied Mathematics,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong, China

Kenneth C. Land

Department of Sociology and Social Science Research Institute,
Duke University,
27708, Durham, NC, USA

(Communicated by the associate editor name)

Abstract. Negative binomial regression has been widely applied in various research settings to account for counts with overdispersion. Yet, when the gamma scale parameter, ν, is parameterized, there is no direct algorithmic solution to the Fisher Information matrix of the associated heterogeneous negative binomial regression, which seriously limits its applications to a wide range of complex problems. In this research, we propose a numerical method to calculate the Fisher information of heterogeneous negative binomial regression and accordingly develop a preliminary framework for analyzing incomplete counts with overdispersion. This method is implemented in R and illustrated using an empirical example of teenage drug use in America.

1. **Introduction.** While negative binomial regression models have often been used to account for count data with overdispersion [1, 13, 14, 22], their applications have been hindered by two critical issues. First, although it is theoretically desirable to compute the expected Fisher information to estimate negative binomial models, empirical/observed Fisher information is nevertheless used in practice for the sake

2010 *Mathematics Subject Classification.* Primary: 62J12; Secondary: 49M15.

Key words and phrases. Regression Analysis, Incomplete Counts, Overdispersion, Heterogeneous Negative Binomial Regression, Fisher Information, Gamma Scale Parameter.

The first author is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 25301115).

* Corresponding author: Qiang Fu.

of computational and arithmetic simplicity [5, 21, 25]. Moreover, this issue becomes more challenging when the gamma scale parameter, ν, is parameterized and the so-called "heterogeneous negative binomial regression" is at stake [16]. In his highly cited book on negative binomial regression, Hilbe notes that there is no software package available for estimating heterogeneous negative binomial regression [16]. Second, despite the fact that incomplete rather than exact counts are being collected in various epidemiological, demographic, and social settings [6, 11, 9, 15], little serious effort has been made to implement negative binomial regression when the counts being studied are incomplete and/or overdispersed. By proposing a numerical method to calculate Fisher information and further exploring its application to incomplete counts, we provide a method/software to estimate (heterogeneous) negative binomial regression that is broadly applicable in empirical research. We particularly consider the application of heterogeneous negative binomial regression to incomplete counts. Our algorithm is implemented in R and results are illustrated by an empirical analysis of data on drug use among American youth.

2. Negative Binomial Distributions and Parameter Estimation.

Let $\mu, \nu \in (0, \infty)$ and consider the negative binomial distribution $\mathsf{NB}(\mu, \nu)$ on the set $\mathbb{N} = \{0, 1, \ldots\}$ of non-negative integers. If $X \sim \mathsf{NB}(\mu, \nu)$, then

$$\omega_k = \omega_k(\mu, \nu) := \text{Prob}(X = k) = \frac{\Gamma(k+\nu)}{k!\Gamma(\nu)} \pi^\nu (1-\pi)^k, \quad k \in \mathbb{N},$$

where $\pi = \frac{\nu}{\mu+\nu}$, and $\Gamma(t) = \int_0^\infty e^{-u} u^{t-1} du$ for $t > 0$ is the Gamma function. Negative binomial distributions are widely used in the social sciences to model count data [3]. In particular, it is well known that for $X \sim \mathsf{NB}(\mu, \nu)$, $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \mu + \frac{\mu^2}{\nu}$, and as $\nu \to \infty$, $\mathsf{NB}(\mu, \nu)$ converges in law to the Poisson distribution with mean μ. There are many other forms of definitions for negative binomial distributions, for example, sometimes people use $\alpha = \frac{1}{\nu}$ to replace the parameter ν. See [16] for a comprehensive review.

Parameter inference for negative binomial distributions is well documented. For example, with a sample $\{X_i\}_{i=1}^n$ drawn independently from $\mathsf{NB}(\mu^*, \nu^*)$, the log-likelihood function takes the form

$$\ell_n(\mu, \nu) = n \log \frac{\nu^\nu}{\Gamma(\nu)(\mu+\nu)^\nu} + \sum_{i=1}^n \log \left[\frac{\Gamma(X_i+\nu)}{X_i!} \left(\frac{\mu}{\nu+\mu} \right)^{X_i} \right]. \tag{1}$$

ℓ_n can be maximized through ordinary optimization methods, for example, gradient ascent or Newton's method. The Fisher information of $\mathsf{NB}(\mu, \nu)$ is given by the 2-by-2 matrix

$$\mathbb{I} = \mathbb{I}(\mu, \nu) = \begin{pmatrix} I_{\mu\mu} & I_{\mu\nu} \\ I_{\nu\mu} & I_{\nu\nu} \end{pmatrix},$$

where $I_{\mu\mu} = \frac{\nu}{\mu(\mu+\nu)}$ and $I_{\mu\nu} = I_{\nu\mu} = 0$, of which we give the detailed derivation in Appendix for the sake of completeness. In particular, it is well known that $I_{\nu\nu}$

does not have a simple form for computation,

$$\begin{aligned} I_{\nu\nu} =& - \mathbb{E}_{X\sim\mathsf{NB}(\mu,\nu)} \frac{\partial^2}{\partial \nu^2} \log \omega_X \\ =& - \mathbb{E}\left[\psi_1(X+\nu) - \psi_1(\nu) + \frac{1}{\nu} - \frac{1}{\nu+\mu} - \frac{\mu - X}{(\mu+\nu)^2}\right] \\ =& \psi_1(\nu) - \frac{\mu}{\nu(\mu+\nu)} - \mathbb{E}[\psi_1(X+\nu)], \end{aligned}$$

where $\psi_1(t) = \frac{d^2}{dt^2}\log\Gamma(t)$ is the trigamma function. Here, the last expectation $\mathbb{E}[\psi_1(X+\nu)]$ has an infinite series expansion. It is surprising that (at least to the best of our knowledge) there is no algorithm that readily computes this expectation for any parameter $\mu, \nu \in (0, \infty)$ with a satisfactory time complexity. Instead, scholars resort to empirical/observed Fisher information to solve this issue [16].

Let $(\hat{\mu}_n, \hat{\nu}_n)$ be the maximum likelihood estimator (MLE) derived from (1) with a sample $\{X_i\}_{i=1}^n$ drawn independently from $\mathsf{NB}(\mu^*, \nu^*)$. It is well understood (for example, see the textbook [23] for a detailed treatment) that as the sample size $n \to \infty$,

$$\sqrt{n}\begin{pmatrix} \hat{\mu}_n - \mu^* \\ \hat{\nu}_n - \nu^* \end{pmatrix} \xrightarrow{Law} N(\mathbf{0}, [\mathbb{I}(\mu^*, \nu^*)]^{-1}), \tag{2}$$

where $\mathbb{I}(\mu^*, \nu^*)$ is the Fisher information of $\mathsf{NB}(\mu^*, \nu^*)$. The asymptotic distribution (2) is used to build the confidence intervals when the sample size n is large. In practice, due to the computational issue, $I_{\nu\nu}$ is usually replaced by the empirical Fisher information (EF)

$$I_{\nu\nu}^{\mathsf{EF}} := \sum_{i=1}^n \left(\frac{\partial}{\partial \nu} \log \omega_{X_i}\right)^2.$$

It has been reported (see, e.g., [21, 5]) that even with a large sample, EF may not be a good approximation of the Fisher information, and the theoretical assumption to guarantee the convergence $I_{\nu\nu}^{\mathsf{EF}} \to I_{\nu\nu}$ may usually be violated in real applications. When $I_{\nu\nu}$ is small, the error $I_{\nu\nu} - I_{\nu\nu}^{\mathsf{EF}}$ can be magnified in the inverse matrix $\mathbb{I}^{-1}$.

In this paper, we use a numerical method to estimate $I_{\nu\nu}$. Since we already have mature algorithms for computing $\psi_1(\nu)$ and $\omega_k(\mu, \nu)$ (for example, `trigamma`(ν) and `dnbinom`$(\mathtt{x} = k, \mathtt{size} = \nu, \mathtt{mu} = \mu)$ in R, respectively), we focus on estimating the expectation of $\psi_1(X+\nu)$. Noting that on $(0, \infty)$, ψ_1 is positive and decreasing, we have the following estimate:

$$0 < f_{\mathsf{L}}^m(\mu, \nu) < \mathbb{E}\left[\psi_1(X+\nu)\right] = \sum_{k=0}^{\infty} \psi_1(k+\nu)\omega_k < f_{\mathsf{L}}^m(\mu, \nu) + f_{\mathsf{D}}^m(\mu, \nu),$$

where for $m \in \mathbb{N}$,

$$f_{\mathsf{L}}^m(\mu, \nu) = \sum_{k=0}^{m} \psi_1(k+\nu)\omega_k, \quad \text{and}$$

$$f_{\mathsf{D}}^m(\mu, \nu) = \psi_1(m+1+\nu) \sum_{k=m+1}^{\infty} \omega_k.$$

It is easy to see that as $m \to \infty$, $f_{\mathsf{L}}^m \uparrow \mathbb{E}\left[\psi_1(X+\nu)\right]$ and $f_{\mathsf{L}}^m + f_{\mathsf{D}}^m \downarrow \mathbb{E}\left[\psi_1(X+\nu)\right]$. The computation of f_{L}^m takes m evaluations of ψ_1 and ω_k each, and m multiplications. The computation of f_{D}^m is of time complexity $O(1)$ since the sum $\sum_{m+1}^{\infty} \omega_k$

is usually implemented (for example, by `pnbinom` in R). We propose the following approximation

$$\mathbb{E}\left[\psi_1(X+\nu)\right] \approx f_{\mathsf{L}}^m(\mu,\nu) + \frac{1}{2} f_{\mathsf{D}}^m(\mu,\nu), \tag{3}$$

of which the error is bounded by $\frac{1}{2} f_{\mathsf{D}}^m(\mu,\nu)$. The relative error is estimated by

$$\frac{\left|\mathbb{E}\left[\psi_1(X+\nu)\right] - (f_{\mathsf{L}}^m(\mu,\nu) + \frac{1}{2} f_{\mathsf{D}}^m(\mu,\nu))\right|}{\mathbb{E}\left[\psi_1(X+\nu)\right]} \leq \frac{f_{\mathsf{D}}^m(\mu,\nu)}{2 f_{\mathsf{L}}^m(\mu,\nu)}.$$

To show the convergence speed of the approximation in (3), we provide a simulation to find an integer m which is sufficiently large to make sure that the relative error is smaller than 10^{-3},

$$\frac{f_{\mathsf{D}}^m(\mu,\nu)}{2 f_{\mathsf{L}}^m(\mu,\nu)} \leq 10^{-3}.$$

For different μ and ν, an estimate of m is plotted in Figure 1. We also assess the m needed to guarantee the relative error 10^{-7}, and present the results in Figure 2. Here, we choose $\mu, \nu \in \{10^{l/2} : l = -12, -11, \cdots, 12\}$. For each pair (μ, ν), the evaluation of f_{L}^m with $m = 10^7$ is further timed on a laptop computer with 3.40GHz CPU and 32GB RAM. The mean computation time is 2.79 seconds with maximum 3.58 seconds. We see that the method we proposed achieves satisfactory precision within an acceptable time period for a wide scope of μ and ν.

For other parameterization of the negative binomial distributions, we consider the one-to-one differentiable change of variables $\alpha = \alpha(\mu,\nu)$ and $\beta = \beta(\mu,\nu)$. Write the Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} \frac{\partial \mu}{\partial \alpha} & \frac{\partial \mu}{\partial \beta} \\ \frac{\partial \nu}{\partial \alpha} & \frac{\partial \nu}{\partial \beta} \end{pmatrix}.$$

Then, the Fisher information matrix with respect to (α, β) is $\mathbb{I}(\alpha,\beta) = \mathbf{J}^T \mathbb{I}(\mu,\nu)\mathbf{J}$. We give the following two examples.

Example 1. Consider $\alpha = \nu$ and $\beta = \mu/\nu$, so $\mu = \alpha\beta$. The probability mass function is

$$\text{Prob}(X = k) = \frac{\Gamma(\alpha+k)}{k!\Gamma(\alpha)} \frac{\beta^k}{(1+\beta)^{\alpha+k}}.$$

We have

$$\mathbf{J} = \begin{pmatrix} \beta & \alpha \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad \mathbb{I}(\alpha,\beta) = \begin{pmatrix} \psi_1(\alpha) - \mathbb{E}[\psi_1(X+\alpha)] & \frac{1}{1+\beta} \\ \frac{1}{1+\beta} & \frac{\alpha}{\beta(1+\beta)} \end{pmatrix}.$$

This model was used in [8].

Example 2. Set $\alpha = \nu$ and $\beta = \frac{\mu}{\mu+\nu}$. We have $\mu = \frac{\alpha\beta}{1-\beta}$. The probability mass function is

$$\text{Prob}(X = k) = \frac{\Gamma(\alpha+k)}{k!\Gamma(\alpha)} (1-\beta)^\alpha \beta^k.$$

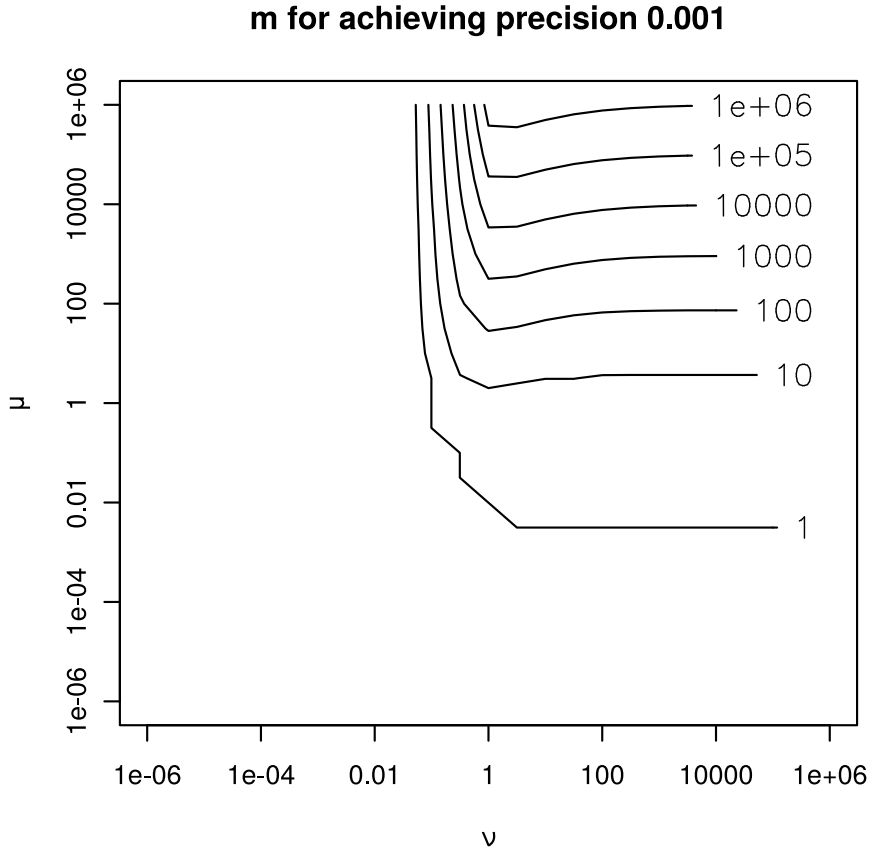

FIGURE 1. The time complexity m for achieving relative error 10^{-3}.

The Jacobian matrix, and the Fisher information matrix are respectively,

$$\mathbf{J} = \begin{pmatrix} \frac{\beta}{1-\beta} & \frac{\alpha}{(1-\beta)^2} \\ 1 & 0 \end{pmatrix}, \quad \text{and}$$

$$\mathbb{I}(\alpha,\beta) = \begin{pmatrix} \psi_1(\alpha) - \mathbb{E}[\psi_1(X+\alpha)] & \frac{1}{1-\beta} \\ \frac{1}{1-\beta} & \frac{\alpha}{\beta(1-\beta)^2} \end{pmatrix}.$$

3. **Parameter Estimation with Grouped and Right-censored (GRC) Data.** In the real applications of parameter inference, observed data are often grouped and right-censored. For example, let $X \sim \mathsf{NB}(\mu,\nu)$, and assume that a questionnaire is designed to record X only by the following five groups: 0, 1-3, 4-6, 7, 8+. Then, the data available to the parameter inference algorithm is only the index of the group X belongs to for each observation. Although by this process of grouping and right-censoring, some information is lost, one obtains some benefit of privacy protection and the relief of interviewee fatigue. Therefore, grouping and right-censoring is a widely adopted practice in data-driven social science research.

Mathematically, we formulate the grouping and right-censoring process as follows. Let $0 < N < \infty$ be the number of groups. We use a sequence $0 = l_1 < l_2 < \cdots < l_{N+1} = \infty$ of integers and infinity, to define the boundaries of different

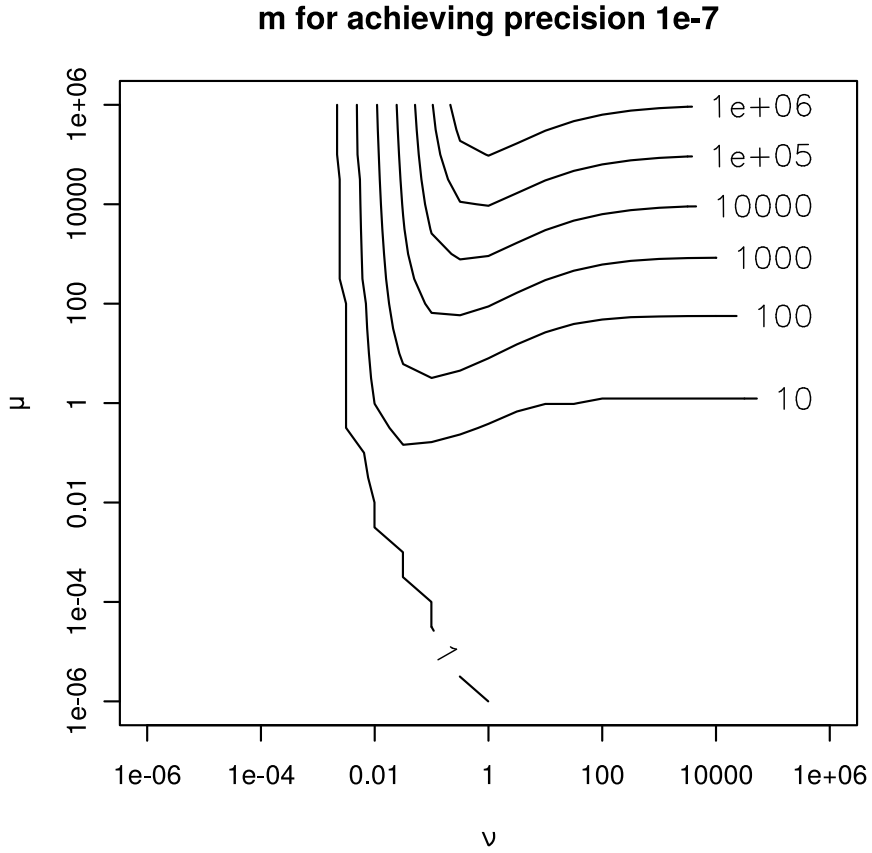

FIGURE 2. The time complexity m for achieving relative error 10^{-7}.

groups. In particular, for $1 \le k \le N$, the k'th group is defined by

$$\text{Group}_k = \{j \in \mathbb{N} : l_k \le j < l_{k+1}\}.$$

Write $\mathcal{G} = \{\text{Group}_k : 1 \le k \le N\}$ for the grouping scheme.

Let $\{X_i\}_{i=1}^n \sim \mathsf{NB}(\mu^*, \nu^*)$ be a sequence of independent random variables. Define the observed data $\{X_i^{\mathcal{G}}\}_{i=1}^n$ by the group indexes. For $1 \le k \le N$,

$$X_i^{\mathcal{G}} = k \quad \text{if and only if } l_k \le X_i < l_{k+1}.$$

The probability of the k'th group is the sum of the values of the probability mass function on the group. Therefore, the log-likelihood function takes the form

$$\ell_n^{\mathcal{G}}(\mu, \nu) = \sum_{i=1}^n \log \theta^{\mathcal{G}}(X_i^{\mathcal{G}}, \mu, \nu) = \sum_{k=1}^N n_k \log \theta_k^{\mathcal{G}}, \tag{4}$$

where for $1 \le k \le N$,

$$\theta_k^{\mathcal{G}} = \theta^{\mathcal{G}}(k, \mu, \nu) = \sum_{j=l_k}^{l_{k+1}-1} \omega_j,$$

is the likelihood of the k'th group, and n_k is the number of the sample points in $\{X_i^{\mathcal{G}}\}_{i=1}^n$ that equals k.

It is easy to see that for $X \sim \mathsf{NB}(\mu, \nu)$, its group index $X^{\mathcal{G}}$ has a multinomial distribution $M^{\mathcal{G}}(\mu, \nu) = M(\theta_k^{\mathcal{G}} : 1 \le k \le N)$ with one trial, on the set $\{1, \dots, N\}$, and

$$\mathrm{Prob}(X^{\mathcal{G}} = k) = \theta^{\mathcal{G}}(k, \mu, \nu) \quad \text{for } 1 \le k \le N.$$

The large sample distribution of the MLE derived from (4) could thus be characterized by the Fisher information matrix $\mathbb{I}^{\mathcal{G}}(\mu, \nu)$ of $M^{\mathcal{G}}(\mu, \nu)$. Here,

$$\mathbb{I}^{\mathcal{G}} = \begin{pmatrix} I^{\mathcal{G}}_{\mu\mu} & I^{\mathcal{G}}_{\mu\nu} \\ I^{\mathcal{G}}_{\nu\mu} & I^{\mathcal{G}}_{\nu\nu} \end{pmatrix}, \tag{5}$$

with

$$I^{\mathcal{G}}_{\mu\mu} = \sum_{k=1}^{N} \frac{1}{\theta_k^{\mathcal{G}}} \left(\frac{\partial}{\partial \mu} \theta_k^{\mathcal{G}} \right)^2, \quad I^{\mathcal{G}}_{\mu\nu} = I^{\mathcal{G}}_{\nu\mu} = \sum_{k=1}^{N} \frac{1}{\theta_k^{\mathcal{G}}} \frac{\partial \theta_k^{\mathcal{G}}}{\partial \mu} \frac{\partial \theta_k^{\mathcal{G}}}{\partial \nu}, \quad \text{and}$$

$$I^{\mathcal{G}}_{\nu\nu} = \sum_{k=1}^{N} \frac{1}{\theta_k^{\mathcal{G}}} \left(\frac{\partial}{\partial \nu} \theta_k^{\mathcal{G}} \right)^2.$$

Intuitively, a finer grouping scheme provides more information for parameter inference. Mathematically, we say that a grouping scheme $\mathcal{G}$ is finer than another grouping scheme $\mathcal{G}'$, if each group in $\mathcal{G}$ is entirely included in only one group in $\mathcal{G}'$. In other words, $\mathcal{G}$ is obtained by cutting one or more groups of $\mathcal{G}'$, into smaller groups. We write $\mathcal{G} \succeq \mathcal{G}'$ when $\mathcal{G}$ is finer than or equal to $\mathcal{G}'$.

For a family of distributions parameterized by a single parameter, the Fisher information matrices have size 1×1, and they can thus be compared as non-negative real numbers. For a general Fisher information matrix of size $r \times r$ with $r \ge 1$, one may use the Loewner partial order (see, e.g., [17, Section 7.7]). Let A and B be symmetric matrices of size $r \times r$. We write $A \succeq B$ if $A - B$ is positive semi-definite.

The following theorem verifies the above intuition that a finer grouping scheme never leads to less Fisher information.

Theorem 3.1. *Let $\mathcal{G}$ and $\mathcal{G}'$ be two grouping schemes of $\mathbb{N}$. Let $\mathbb{I}^{\mathcal{G}}$ and $\mathbb{I}^{\mathcal{G}'}$ be the associated Fisher information we defined in (5), respectively. If $\mathcal{G} \succeq \mathcal{G}'$, then $\mathbb{I}^{\mathcal{G}}(\mu, \nu) \succeq \mathbb{I}^{\mathcal{G}'}(\mu, \nu)$ for any $\mu, \nu \in (0, \infty)$.*

Theorem 3.1 is formulated for GRC data from negative binomial distributions. Similar theorems for the GRC data from Poisson and zero-inflated Poisson distributions are obtained in [10]. Nevertheless, we find that these theorems can easily be extended to a more general setting. We now formulate the setting, give the general theorem and its proof, and then prove Theorem 3.1 as a corollary.

Let S be a measurable space, on which we define a family of probability distributions $\{\rho(x|\boldsymbol{\mu} = (\mu_1, \dots, \mu_r))|\boldsymbol{\mu} \in \Upsilon\}$ parameterized on an open set $\Upsilon \subset \mathbb{R}^r$. We group several technical assumptions into the following definition.

Definition 3.2 (feasible partition)**.** A finite family $\mathcal{G} = \{S_i : 1 \le i \le N, S_i \subset S\}$ of subsets of S is referred to as a feasible partition of S if the following conditions are all satisfied.

1. S_i's form a partition of S. That is, $S_i \cap S_j = \emptyset$ when $i \ne j$, and $\cup_{i=1}^N S_i = S$.
2. For each $1 \le i \le N$, S_i is measurable, and $\int_{S_i} d\rho(x|\boldsymbol{\mu}) > 0$ for any $\boldsymbol{\mu} \in \Upsilon$.

3. For any $1 \le i \le N$, the integral $\int_{S_i} d\rho(x|\boldsymbol{\mu})$ is a continuously differentiable function of $\boldsymbol{\mu}$ on Υ.

The existence of feasible partitions, while not necessarily trivial, is assumed.

The notion of a feasible partition is a generalization of the grouping scheme we study in this paper. For two feasible partitions $\mathcal{G}$ and $\mathcal{G}'$, we say $\mathcal{G}$ is finer than $\mathcal{G}'$ and write $\mathcal{G} \succeq \mathcal{G}'$, if each set in $\mathcal{G}$ is entirely contained in only one set of $\mathcal{G}'$.

For a feasible partition $\mathcal{G} = \{S_i : 1 \le i \le N\}$ of S, write $M^{\mathcal{G}} = M(\theta_i^{\mathcal{G}}(\boldsymbol{\mu}) : 1 \le i \le N)$ the multinomial distribution on $\{1, \ldots, N\}$ with one trial, where $\theta_i^{\mathcal{G}} = \int_{S_i} d\rho(x|\boldsymbol{\mu})$. Let $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\mu}) \in \mathbb{R}^{r\times r}$ be the Fisher information of $M^{\mathcal{G}}$. For another feasible partition $\mathcal{G}'$, define $\mathbb{I}^{\mathcal{G}'}(\boldsymbol{\mu})$ in the same way above by substituting $\mathcal{G}$ with $\mathcal{G}'$.

Theorem 3.3. *Let $\mathcal{G} \succeq \mathcal{G}'$ be two feasible partitions of S. We have*

$$\mathbb{I}^{\mathcal{G}}(\boldsymbol{\mu}) \succeq \mathbb{I}^{\mathcal{G}'}(\boldsymbol{\mu}), \quad \text{for any } \boldsymbol{\mu} \in \Upsilon. \tag{6}$$

Furthermore,

$$\operatorname{rank}(\mathbb{I}^{\mathcal{G}}(\boldsymbol{\mu}) - \mathbb{I}^{\mathcal{G}'}(\boldsymbol{\mu})) \le |\mathcal{G}| - |\mathcal{G}'|. \tag{7}$$

The following theorem gives some insights on the structure of $\mathbb{I}^{\mathcal{G}}(\boldsymbol{\mu}) - \mathbb{I}^{\mathcal{G}'}(\boldsymbol{\mu})$.

Theorem 3.4. *Let $\mathcal{G} \succeq \mathcal{G}'$ be two feasible partitions of S, such that $|\mathcal{G}| = |\mathcal{G}'| + 1$, and the union of the sets A and B in $\mathcal{G}$ forms a set $A \cup B$ in $\mathcal{G}'$. Let $w^A = w^A(\boldsymbol{\mu}) = \int_A d\rho(x|\boldsymbol{\mu})$ denote the probability of the set A under the distribution $\rho(\cdot|\boldsymbol{\mu})$, as a function of $\boldsymbol{\mu}$. For any function $f(\boldsymbol{\mu})$ of $\boldsymbol{\mu}$, denote ∇f the gradient, which is a vector-valued function of $\boldsymbol{\mu}$. One has*

$$\mathbb{I}^{\mathcal{G}} - \mathbb{I}^{\mathcal{G}'} = \frac{(w^B)^3}{w^A(w^A + w^B)} \left(\nabla \frac{w^A}{w^B}\right) \left(\nabla \frac{w^A}{w^B}\right)^T. \tag{8}$$

Therefore, $\mathbb{I}^{\mathcal{G}} = \mathbb{I}^{\mathcal{G}'}$ only at the stationary points of w^A/w^B, i.e., the points where $\nabla(w^A/w^B) = \mathbf{0}$.

Proof. For $1 \le i \le r$, define $w_i^A = \partial w^A/\partial \mu_i$ the partial derivative of w^A with respect to the i'th coordinate μ_i of $\boldsymbol{\mu}$, and define w_i^B in the same way. For $1 \le i, j \le r$, the (i, j) entry of $\mathbb{I}^{\mathcal{G}} - \mathbb{I}^{\mathcal{G}'}$ is

$$\begin{aligned}
\left[\mathbb{I}^{\mathcal{G}} - \mathbb{I}^{\mathcal{G}'}\right]_{i,j} =& \frac{w_i^A w_j^A}{w^A} + \frac{w_i^B w_j^B}{w^B} - \frac{(w_i^A + w_i^B)(w_j^A + w_j^B)}{w^A + w^B} \\
=& \frac{(w^B)^2 w_i^A w_j^A + (w^A)^2 w_i^B w_j^B - w^A w^B (w_i^A w_j^B + w_i^B w_j^A)}{w^A w^B (w^A + w^B)} \\
=& \frac{(w^B w_i^A - w^A w_i^B)(w^B w_j^A - w^A w_j^B)}{w^A w^B (w^A + w^B)} \\
=& \frac{(w^B)^3}{w^A(w^A + w^B)} \left(\frac{\partial}{\partial \mu_i} \frac{w^A}{w^B}\right) \left(\frac{\partial}{\partial \mu_j} \frac{w^A}{w^B}\right).
\end{aligned}$$

The proof is completed. □

Proof of Theorem 3.3. The relation (6) is a direct corollary of Theorem 3.4 by observing that the matrix in the right-hand side of (8) is positive semi-definite. For proving (7), suppose $A_k \in \mathcal{G}$ for $1 \le k \le t$ and $\cup_{k=1}^t A_k \in \mathcal{G}'$. Write

$u_k = \int_{A_k} d\rho(x|\boldsymbol{\mu})$. We apply Theorem 3.4 $t-1$ times to obtain that the splitting of $\cup_{k=1}^{t} A_k \in \mathcal{G}'$ to $\{A_k : 1 \le k \le t\}$ contributes the following terms to the information matrix difference $\mathbb{I}^{\mathcal{G}} - \mathbb{I}^{\mathcal{G}'}$,

$$\sum_{k=1}^{t-1} \frac{(u_{k+1})^3}{u_1^k u_1^{k+1}} \left(\nabla \frac{u_1^k}{u_{k+1}}\right) \left(\nabla \frac{u_1^k}{u_{k+1}}\right)^T,$$

where $u_1^l := u_1 + u_2 + \cdots + u_l$ for $1 \le l \le t$. The proof is thus completed. □

Proof of Theorem 3.1. We need only to verify that any grouping scheme we described for $\mathbb{N}$ in this paper is feasible in the sense of Definition 3.2. We check the three items in Definition 3.2 one by one. First, any grouping scheme is a partition and satisfies Item 1. Second, for any $\mu, \nu \in (0, \infty)$, any subset of $\mathbb{N}$ is measurable with respect to $\mathsf{NB}(\mu, \nu)$, and any integer $k \in \mathbb{N}$ carries positive probability, so Item 2 is satisfied. Item 3 follows from the fact that for any $k \in \mathbb{N}$, the probability mass function of $\mathsf{NB}(\mu, \nu)$ is analytic with respect to both μ and ν. □

Remark 1. Here we require that both partitions $\mathcal{G}$ and $\mathcal{G}'$ are finite: $|\mathcal{G}|, |\mathcal{G}'| < \infty$. This requirement can be relaxed to that one or both of the partitions are countable. Nevertheless, considering only finite partitions is sufficient for the study of GRC data, and helps us to save the treatment of the countable partition as a limit of finite partitions.

In summary, Theorem 3.1 shows that a finer grouping scheme can never leads to less Fisher information. Theorems 3.3 and 3.4 further demonstrate how Fisher information increases as a grouping scheme is divided finer. In particular, when one group $A \cup B$ in a grouping scheme is divided into two, A and B, the Fisher information matrix *increases* by a rank-one matrix of which the corresponding eigenvector parallels the gradient of w^A/w^B. These theorems lend support to search for the optimal grouping scheme in two aspects. First, with the constraint that the maximum possible number of groups is N, the optimizer can only exist among these grouping schemes with N groups. Second, in the process of search, a grouping scheme containing $k > N$ groups yet showing a suboptimal score could be used to assess/purge all grouping schemes coarser than this specific scheme. While the first aspect has been adopted in this research, the second one remains an interesting topic for future research.

4. **Data, Variables and Results.** As an empirical illustration of the estimation methods developed in the previous section, we use data from the Monitoring the Future (MTF) project, which is a nationally representative survey administered by The Institute for Social Research in The University of Michigan [19, 12]. Since the year 1975, the MTF project annually investigates a variety of behaviors and attitudes related to social norms, violence victimization, juvenile delinquency and drug use among American youth. Only 12^{th} graders were included in the initial waves of the survey with $8^{\text{th}}/10^{\text{th}}$ graders added to the survey since the year 1991. The dataset being analyzed here is retrieved from the 2012 wave of the survey and students from all three (8^{th}, 10^{th} and 12^{th}) grades are included. All observations containing missing values are listwise deleted and the total sample size for analysis is 8,874.

The outcome variable of the heterogeneous negative binomial regression analysis is respondents' lifetime marijuana use. Adolescents participating in the survey were

asked how many occasions they used marijuana in their lifetime and the response categories were "0 occasion", "1-2 times", "3-5 times", "6-9 times", "10-19 times", "20-39 times", and "40 and above times". Independent variables of this regression analysis consist of five demographic variables: 10^{th} *graders* and 12^{th} *graders* are two dummy variables (8^{th} graders as reference) denoting the age/grade of corresponding respondents; *Male* is coded as 1 if a respondent is male and 0 otherwise; *African American* is another variable denoting the racial/ethnic group of a respondent and it is coded as 1 if a respondent is Black; finally, *Metropolitan area* is coded as 1 if a respondent grew up in a metropolitan area in the United States and coded as 0 if s/he was from a small town or other rural areas. All these five covariates are used to model μ and ν, respectively, via a log-link function [16]. We wrote a package in R, GRCData, to implement the heterogeneous negative binomial regression analysis.

Results of our empirical analysis are shown in Table 1. As compared to students from the 8^{th} grade, both 10^{th} and 12^{th} graders are significantly and positively associated with μ, or the frequency of marijuana use. This strong positive association between age and marijuana use among adolescents in America is consistent with existing literature [18]. Also, male students show significantly higher frequencies of marijuana use as compared to their female counterparts. Living in metropolitan areas is not significantly associated with the frequency of marijuana use. One interesting finding is that African Americans are less likely to use marijuana. While fundamental causes for the racial disparity in marijuana use warrant more in-depth investigation, several conclusions from existing studies may help us understand why African American adolescents exhibit lower levels of marijuana use [7, 24, 20]. First, a racial crossover probably exists such that black students reported more marijuana use than did white students in lower grades (e.g., grade 9), but a reversed pattern is observed in higher school grades [7]. Second, the earlier years of the MTF data African American students showed lower use of any illicit drug (including marijuana) than did Whites and Hispanics, but the gap has narrowed in recent years [18]. Third, mixed findings pertaining to racial disparities in marijuana use have also been reported, which suggests that marijuana use might be more affected by socio-economic status or class rather than by race/ethnicity [24, 7]. Next, we study the variance of marijuana use. Again, due to the definition of ν, a positive association with ν suggests a smaller variance in marijuana use for a specific demographic group. As suggested by the second panel in Table 1, 10^{th} graders, 12^{th} graders, and African Americans are all positively associated with ν and thus have lower variances of marijuana use. The associations between ν and the other two covariates (sex and living in metropolitan areas) are not significant. Various measures of goodness of fit including the Akaike information criterion, the Bayesian Information criterion, and the McFadden's R^2, are also reported for readers' information [2, 4].

5. **Conclusions.** In this research, we develop a new numerical method to calculate the (expected) Fisher information associated with the (heterogeneous) negative binomial regression, and the application of this method to a special case of count data, namely, grouped and right censored counts. Based on a new package developed in R, the application is further illustrated using an empirical example of drug use among American youth. It should be noted that the numerical method introduced here could serve as a general tool for the estimation of (heterogeneous) negative binomial regression models in empirical analyses.

	Coefficient	Standard error	Z value	95% confidence interval
Covariates for estimating μ				
Intercept	0.677 ***	0.183	3.696	[0.318, 1.036]
10th graders	1.551 ***	0.153	10.145	[1.251, 1.850]
12th graders	2.002 ***	0.168	11.927	[1.673, 2.331]
Male	1.268 ***	0.125	10.143	[1.023, 1.513]
African American	-0.796 ***	0.149	-5.361	[-1.087, -0.505]
Metropolitan areas	0.148	0.150	0.983	[-0.147, 0.442]
Covariates for estimating ν				
Intercept	-3.627 ***	0.082	-44.331	[-3.787, -3.466]
10th graders	0.972 ***	0.068	14.374	[0.839, 1.104]
12th graders	1.332 ***	0.074	18.018	[1.188, 1.477]
Male	-0.006	0.051	-0.107	[-0.106, 0.095]
African American	0.268 ***	0.077	3.480	[0.117, 0.418]
Metropolitan areas	0.117 .	0.063	1.844	[-0.007, 0.240]
Goodness of fit				
AIC	18400		BIC	18480
McFadden's R2	0.04828	McFadden's adjusted R2		0.04703

Note: *** p<0.001 ** p<0.01 * p<0.05 . P<0.1

TABLE 1. Heterogeneous negative-binomial regression analysis of lifetime marijuana use among American youth (Number of observations=8,874). Data source: the 2012 wave of the Monitoring the Future study.

Appendix. Here we give the detailed computation of $I_{\mu\mu}$, $I_{\mu\nu}$, and $I_{\nu\mu}$ of $\mathsf{NB}(\mu,\nu)$ for completeness. First,

$$\begin{aligned}\frac{\partial^2}{\partial\mu^2}\log\omega_k =& \frac{\partial^2}{\partial\mu^2}\left[\log\frac{\Gamma(k+\nu)\nu^\nu}{k!\Gamma(\nu)} + k\log\mu - (\nu+k)\log(\mu+\nu)\right]\\ =& -\frac{k}{\mu^2} + \frac{\nu+k}{(\mu+\nu)^2}.\end{aligned}$$

So,

$$I_{\mu\mu} = -\mathbb{E}\frac{\partial^2}{\partial\mu^2}\log\omega_X = \frac{1}{\mu} - \frac{1}{\mu+\nu} = \frac{\nu}{\mu(\mu+\nu)}.$$

Meanwhile,

$$\frac{\partial^2}{\partial\nu\partial\mu}\log\omega_k = \frac{\partial}{\partial\nu}\left[\frac{k}{\mu} - \frac{\nu+k}{\mu+\nu}\right] = \frac{k-\mu}{(\mu+\nu)^2},$$

so $I_{\mu\nu} = I_{\nu\mu} = -\mathbb{E}\frac{\partial^2}{\partial\nu\partial\mu}\log\omega_X = 0$.

REFERENCES

[1] P. D. Allison and R. P. Waterman, Fixed–effects negative binomial regression models, *Sociological Methodology*, **32** (2002), 247–265.

[2] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens and J.-S. S. White, Generalized linear mixed models: a practical guide for ecology and evolution, *Trends in Ecology & Evolution*, **24** (2009), 127–135.

[3] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, vol. 53, Cambridge University Press, 2013.

[4] A. C. Cameron and F. A. Windmeijer, R-squared measures for count data regression models with applications to health-care utilization, *Journal of Business & Economic Statistics*, **14** (1996), 209–220.
[5] B. Efron and D. V. Hinkley, Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information, *Biometrika*, **65** (1978), 457–487, URL https://doi.org/10.1093/biomet/65.3.457.
[6] S. Ehsan Saffari, R. Adnan and W. Greene, Hurdle negative binomial regression model with right censored count data, *SORT: Statistics and Operations Research Transactions*, **36** (2012), 0181–194.
[7] K. V. Finn, Patterns of alcohol and marijuana use at school, *Journal of Research on Adolescence*, **16** (2006), 69–77.
[8] R. A. Fisher, The negative binomial distribution, *Annals of Eugenics*, **11** (1941), 182–187.
[9] Q. Fu, X. Guo and K. C. Land, A Poisson-multinomial mixture approach to grouped and right-censored counts, *Communications in Statistics-Theory and Methods*, **47** (2018), 427–447.
[10] Q. Fu, X. Guo and K. C. Land, Optimizing count responses in surveys: A machine-learning approach, *Sociological Methods & Research*, **DOI:10.1177/0049124117747302**, URL https://doi.org/10.1177/0049124117747302.
[11] Q. Fu, K. C. Land and V. L. Lamb, Bullying victimization, socioeconomic status and behavioral characteristics of 12th graders in the united states, 1989 to 2009: Repetitive trends and persistent risk differentials, *Child Indicators Research*, **6** (2013), 1–21, URL https://doi.org/10.1007/s12187-012-9152-8.
[12] Q. Fu, K. C. Land and V. L. Lamb, Violent physical bullying victimization at school: has there been a recent increase in exposure or intensity? an age-period-cohort analysis in the united states, 1991 to 2012, *Child Indicators Research*, **9** (2016), 485–513.
[13] Q. Fu, C. Wu, H. Liu, Z. Shi and J. Gu, Live like mosquitoes: Hukou, rural–urban disparity, and depression, *Chinese Journal of Sociology*, **4** (2018), 56–78.
[14] W. H. Greene, Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, *NYU working paper no. EC-94-10*.
[15] R. M. Groves, F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau, *Survey Methodology*, vol. 561, John Wiley & Sons, 2011.
[16] J. M. Hilbe, *Negative binomial regression*, 2nd edition, Cambridge University Press, Cambridge, 2011, URL https://doi.org/10.1017/CBO9780511973420.
[17] R. A. Horn and C. R. Johnson, *Matrix analysis*, 2nd edition, Cambridge University Press, Cambridge, 2013.
[18] L. D. Johnston, P. M. O'Malley and J. G. Bachman, Monitoring the Future: National results on adolescent drug use: Overview of key findings, *Focus*, **1** (2003), 213–234.
[19] L. D. Johnston, P. M. O'Malley, R. A. Miech, J. G. Bachman and J. E. Schulenberg, Monitoring the future national survey results on drug use, 1975–2016: Overview, key findings on adolescent drug use, https://files.eric.ed.gov/fulltext/ED578534.pdf, 2017, Accessed July 17, 2019.
[20] L. D. Johnston, P. M. O'Malley, R. A. Miech, J. G. Bachman and J. E. Schulenberg, Monitoring the Future national survey results on drug use, 1975-2016: Overview, key findings on adolescent drug use., *Institute for Social Research*.
[21] F. Kunstner, L. Balles and P. Hennig, Limitations of the empirical Fisher approximation, *arXiv preprint arXiv:1905.12558*.
[22] K. C. Land, P. L. McCall and D. S. Nagin, A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models: With empirical applications to criminal careers data, *Sociological Methods & Research*, **24** (1996), 387–442.
[23] E. L. Lehmann and G. Casella, *Theory of point estimation*, 2nd edition, Springer Texts in Statistics, Springer-Verlag, New York, 1998.
[24] L. R. Pacek, R. J. Malcolm and S. S. Martins, Race/ethnicity differences between alcohol, marijuana, and co-occurring alcohol and marijuana use disorders and their association with public health and social problems using a national sample, *The American Journal on Addictions*, **21** (2012), 435–444.
[25] W. W. Piegorsch, Maximum likelihood estimation for the negative binomial dispersion parameter, *Biometrics*, **46** (1990), 863–867.

Received xxxx 20xx; revised xxxx 20xx.

E-mail address: x.guo@polyu.edu.hk
E-mail address: qiang.fu@ubc.ca
E-mail address: yue1995.wang@connect.polyu.hk
E-mail address: kland@soc.duke.edu