A Comparison of Zero-Inflated Models for Modern Biomedical Data

Max Beveridge[a], *Zach Goldstein*[b], *& Hee Cheol Chung*[c]
[a] *Department of Statistics, The George Washington University, Washington, DC*
[b] *Department of Mathematics, Yeshiva University, New York, NY*
[c] *Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC*

https://doi.org/10.33697/ajur.2025.141

Students: maxbeveridge03@gwu.edu, zgoldst3@mail.yu.edu
Mentor: hchung13@charlotte.edu

ABSTRACT
There has been a growing number of datasets exhibiting an excess of zero values that cannot be adequately modeled using standard probability distributions. For example, microbiome data and single-cell RNA sequencing data consist of count measurements in which the proportion of zeros exceeds what can be captured by standard distributions such as the Poisson or negative binomial, while also requiring appropriate modeling of the nonzero counts. Several models have been proposed to address zero-inflated datasets including the zero-inflated negative binomial, hurdle negative binomial model, and the truncated latent Gaussian copula model. This study aims to compare various models and determine which one performs optimally under different conditions using both simulation studies and real data analyses. We are particularly interested in investigating how dependence among the variables, level of zero-inflation or deflation, and variance of the data affects model selection.

KEYWORDS
Zero-Inflated Models; Hurdle Models; Truncated Latent Gaussian Copula Model; Microbiome Data; Gene-Sequencing Data; Zero-Inflation, Negative Binomial; Zero-Deflation

INTRODUCTION

Zero-inflated data refers to datasets with an excess of zeros, where the proportion of zeros cannot be adequately captured by standard probability distributions. Such data frequently arise in various fields, such as health and epidemiology, where large numbers of zeros are often encountered. For example, in substance abuse research, the majority of individuals do not engage in substance abuse, leading to a predominance of zero observations.[1] Similarly, zero-inflated data are common in biomedical research including microbiome studies and single-cell RNA sequencing, where zeros occur due to limited sequencing depth.[2,3] Given the widespread occurrence of zero-inflated data across numerous disciplines, it is essential to model these datasets accurately to ensure valid analyses. Failure to properly account for zero inflation can lead to poor estimation and the potential oversight of statistically significant findings. Accurate modeling of zero-inflated data not only improves the estimation of key parameters but also reduces bias and enhances the understanding of dependence structures.[4] Violating distributional assumptions of statistical tests is one of the "seven deadly sins" of comparative analysis.[5] The consequences of which are biased or incorrect parameter estimates and incorrect p-values. With regard to zero-inflated data, several studies have found that misspecifying the distribution of a general linear model (GLM) when data is zero-inflated leads to invalid statistical inference (e.g., using a Poisson or negative binomial (NB) regression model when the data follows a zero-inflated Poisson or zero-inflated NB distribution).[6]

Zero-inflated models, including zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), hurdle Poisson (HP), and hurdle negative binomial (HNB), have been widely used to model zero-inflated data across fields such as ecology, environmental science (e.g., species counts), economics (e.g., consumer purchases), insurance (e.g., claims data), and criminology (e.g., crime counts in different areas). The key difference between zero-inflated and hurdle models lies in how they handle the excess number of zeros. Zero-inflated models combine a point mass at zero with a standard distribution that also allows non-zero probability at zero. The point mass accounts for structural zeros (inherent zeros), while the non-zero probability from the standard distribution models sampling zeros (zeros that occur by chance). In contrast, hurdle models only account for structural zeros by using a mixture of a point mass at zero

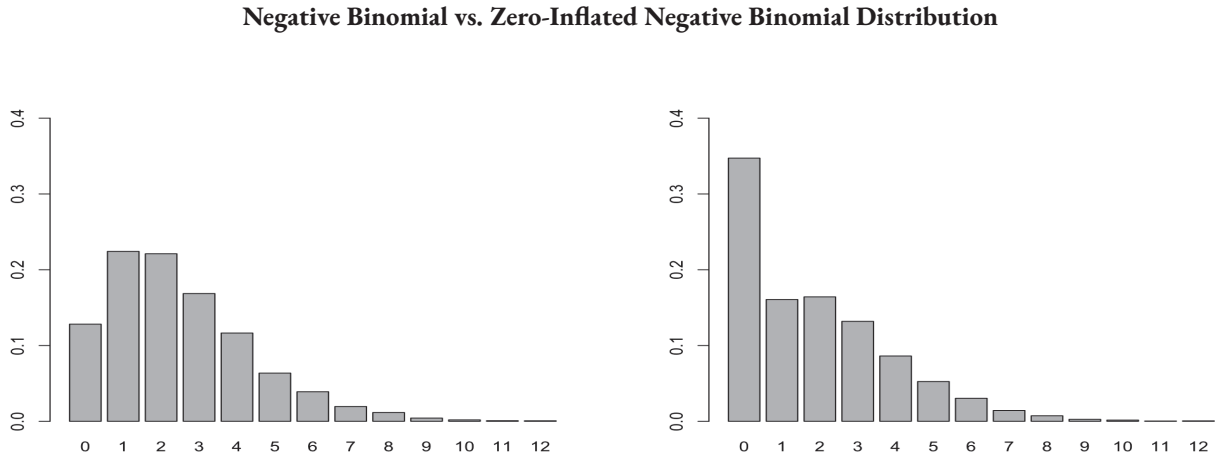

Figure 1. Shown on the left is a standard negative binomial distribution (where $\mu = 2.5$, $r = 5$, and the probability that $Y = 0$ is 0.1317) and on the right is a zero-inflated negative binomial distribution (where $\mu = 2.5$, $r = 5$, and $\pi_Z = 0.25$, and the probability that $Y = 0$ is 0.3487).

In Method and Procedures, we detail each of the zero-inflated models and define terms used in our simulation studies and real data analyses. Then, in Simulation Setting One, we discuss the procedure, results and discussion of comparing the ZINB and HNB models. In Simulation Setting Two and Simulation Setting Three, we discuss the procedure, results, and interpretation of comparing the HNB and TLNPN models (under HNB and TLNPN population data respectively). We then discuss our methods, results, and interpretation of comparing the HNB and TLNPN models using real-world biomedical data in Real Data Analyses. Finally, we will summarize the findings, limitations, and directions for future research in the Conclusion section.

METHOD AND PROCEDURES

In this section, we first detail the zero-inflated models that we will be investigating including the ZINB, HNB, and TLNPN models. We then review important definitions for the proceeding simulation studies and real data analyses.

Models for Zero-Inflated Data

Zero-inflated models account for an excess number of zeros by adjusting the probability of observing zero of a standard probability distribution. In particular, the form of the probability mass function (pmf) of a zero-inflated model is given by:

$$P(Y = y) = \begin{cases} \pi_Z + (1 - \pi_Z)p(y = 0; \mu) & \text{for } y = 0, \\ (1 - \pi_Z)p(y; \mu) & \text{if } y > 0, \end{cases}$$

where $p(\cdot)$ is a pmf of a discrete random variable following a standard distribution, e.g., Poisson or negative binomial distribution, μ is the mean of the distribution, and π_Z is the weight parameter controlling the degree of zero inflation. One of the popularly used zero-inflated models is the zero-inflated negative binomial (ZINB) model with pmf:

$$P(Y = y) = \begin{cases} \pi_Z + (1 - \pi_Z)(\frac{r}{\mu+r})^r & \text{for } y = 0, \\ (1 - \pi_Z)\frac{\Gamma(y+r)}{\Gamma(r)y!}(\frac{\mu}{\mu+r})^y(\frac{r}{\mu+r})^r & \text{if } y > 0, \end{cases}$$

where μ is the mean of the negative binomial model, r is the dispersion parameter, and π_Z is the probability of structural zeros. In zero-inflated models, the probability of observing a zero is given by $\pi_Z + (1 - \pi_Z)p(y = 0; \mu)$. As a result, the probability is bounded below by $p(y = 0; \mu)$, which corresponds to the probability under the standard negative binomial model. Consequently, zero-inflated models cannot account for zero deflation. An illustrative example of the zero-inflated negative binomial distribution is provided in **Figure 1**. In contrast to zero-inflated models, hurdle models are able to account for zero-inflation and zero-deflation.

Hurdle models are distinct from zero-inflated models because they only account for structural zeros and are able to model zero-deflation. Zero-deflation occurs when there are less zero values present than a standard probability distribution would predict. The form of the

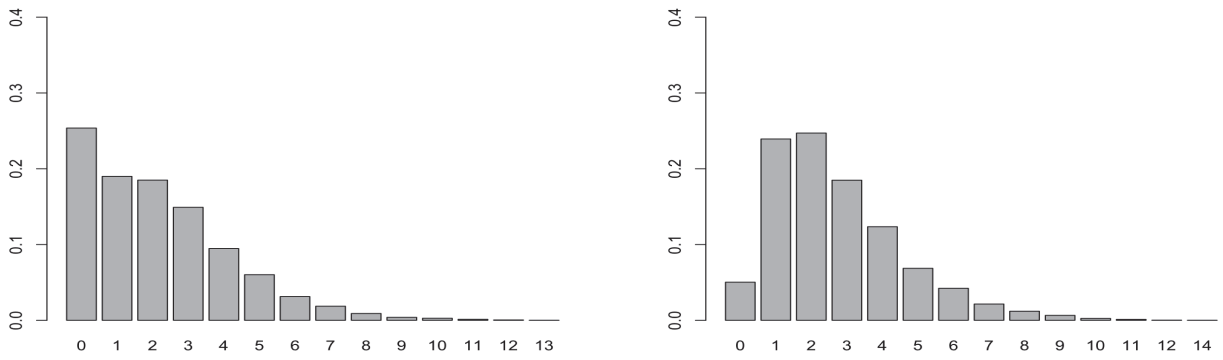

Figure 2. Shown are hurdle negative binomial distributions with $\mu = 2.5$ and $r = 5$. In the left histogram, $\pi_H = 0.25$, so the probability $Y = 0$ is 0.25, and in the right histogram, $\pi_H = 0.05$, so the probability that $Y = 0$ is 0.05. Under a standard negative binomial distribution, the probability that $Y = 0$ is 0.1317.

pmf of a hurdle model is:

$$P(Y = y) = \begin{cases} \pi_H & \text{for } y = 0 \\ (1 - \pi_H)\frac{p(y;\mu)}{1-p(y=0;\mu)} & \text{if } y > 0 \end{cases}$$

where $p(\cdot;\mu)$ is the pmf of a Poisson or negative binomial distribution with mean μ. The parameter π_H is the probability that a structural zero occurs and can take any value from 0 to 1. The hurdle negative binomial (HNB) model is given by:

$$P(Y = y) = \begin{cases} \pi_H & \text{for } y = 0 \\ \frac{1-\pi_H}{1-(\frac{r}{\mu+r})^r}\frac{\Gamma(y+r)}{\Gamma(r)y!}(\frac{\mu}{\mu+r})^y(\frac{r}{\mu+r})^r & \text{if } y > 0. \end{cases}$$

Under the hurdle model, zero occurs with probability π_H, which can be smaller than the probability of $Y = 0$ under the negative binomial model and thus capable of modeling zero-deflated variables. Examples of the hurdle negative binomial distribution can be seen in **Figure 2**. When data involves multiple zero-inflated variables, their associations can be modeled within the generalized linear model framework, assuming covariates are available.

For multiple zero-inflated random variables $Y_1, \ldots, Y_p$, given the covariates $\boldsymbol{x} = (x_1, \ldots, x_{q_1})^\top$ and $\boldsymbol{z} = (z_1, \ldots, z_{q_2})^\top$, which are shared across $Y_1, \ldots, Y_p$, their associations can be modeled within the generalized linear model (GLM) framework. In particular, the ZINB regression model is given by

$$\ln(\mu_j) = \boldsymbol{x}^T\boldsymbol{\beta}_j, \text{ and } \operatorname{logit}(\pi_{Z,j}) = \boldsymbol{z}^T\boldsymbol{\gamma}_j \qquad \textbf{Equation 1.}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{q_1}$ and $\boldsymbol{\gamma}_j \in \mathbb{R}^{q_2}$ are the regression coefficients for the mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ and $\boldsymbol{\pi}_Z = (\pi_{Z,1}, \ldots, \pi_{Z,p})^\top$, respectively, and $\operatorname{logit}(\pi_Z) = \ln\{\pi_Z/(1 - \pi_Z)\}$. For each $j = 1, \ldots, p$, the parameters $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$, and the dispersion parameter r_j, can be estimated using the maximum likelihood estimator. Let $Y_{i1}, \ldots, Y_{ip}, i = 1, \ldots, n$, be a random sample. The log-likelihood function of the jth variable, $L_{ZI,j}$ is defined as $L_{ZI,j} = L_{1,j} + L_{2,j} + L_{3,j} - L_{4,j}$, where

$$L_{1,j} = \sum_{i:y_i=0} \ln\left\{e^{\boldsymbol{z_i}^T\boldsymbol{\gamma}_j} + \left(1 + \frac{\mu_{ij}}{r_j}\right)^{-r_j}\right\}, \quad L_{2,j} = \sum_{i:y_{ij}>0}\sum_{t=0}^{y_{ij}-1} \ln(t + r_j)$$

$$L_{3,j} = \sum_{i:y_i>0} \left\{-\ln(y_{ij}!) - (y_{ij} + r_j)\ln\left(1 + \frac{\mu_{ij}}{r_j}\right) + y_{ij}\ln(r_j^{-1}) + y_{ij}\ln(\mu_{ij})\right\}$$

$$L_{4,j} = \sum_{i=1}^{n} \ln(1 + e^{\boldsymbol{z_i}^T\boldsymbol{\gamma}_j}).$$

Thus, the log-likelihood function of the joint model is given by $L_{ZI} = \sum_j L_{ZI,j}$. The hurdle negative binomial regression model is given by

$$\ln(\mu_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j, \text{ and } \operatorname{logit}(\pi_{H,ij}) = \boldsymbol{x}_i^T \boldsymbol{\gamma}_j \quad \textbf{Equation 2.}$$

The log-likelihood function of the jth variable, $L_{\mathrm{H,j}}$, is given as

$$L_{H,j} = \sum_{i=1}^{n} (I_{y_{ij}=0} \ln(\pi_{H,ij}) + I_{y_i>0}(\ln(1-\pi_{H,ij}) + \ln(h(y_{ij};\mu_{ij},r_j) - \ln(1-(1+r_j\mu_{ij})^{-r_j}))$$

where $h(y_{ij};\mu_{ij},r_j)$ denotes the pmf of the negative binomial distribution with mean μ_{ij} and dispersion parameter r_j.[12] Therefore, the log-likelihood function of the joint hurdle model is given by $L_H = \sum_j L_{H,j}$.

Nevertheless, in real-world applications, such covariates are often not readily available. In these cases, we can only fit the intercept parameters β_0 and γ_0, assuming that all variables are mutually independent. The Gaussian copula model addresses this limitation by utilizing a rank-based correlation estimator. The Gaussian copula model assumes that, for a random vector $\boldsymbol{y} = (Y_1, \ldots, Y_p)^\top$, there exist strictly increasing functions, $g_1, \ldots, g_p$, such that $\boldsymbol{z} = (g_1(Y_1), \ldots, g_p(Y_p))^\top \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for some mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It is important to note that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are not identifiable because, for any constants a_j and b_j, the Gaussian copula model still holds with $g_j^* = a_j + b_j g_j$, $j = 1, \ldots, p$, i.e., $(g_1^*(Y_1), \ldots, g_p^*(Y_p))^\top$ follows $N_p(\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{B\Sigma B})$, where $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ and $\boldsymbol{B} = \operatorname{diag}\{b_j\}_{j=1}^p$. The identifiability issue is commonly addressed by assuming that $\boldsymbol{\mu} = \mathbf{0}_p$ and $\boldsymbol{\Sigma}$ is a positive definite correlation matrix.[11,13] If g_js are differentiable, then we have an analytic expression as $g_j = \Phi^{-1} \circ F_j$, where F_j and Φ^{-1} are the distribution functions of Y_j and standard Gaussian. The Gaussian copula models are often denoted as $\boldsymbol{y} \sim NPN(\mathbf{0}_p, \boldsymbol{\Sigma}, \boldsymbol{g})$.[14]

The Gaussian copula models assume that Y_j are continuous and are thus not valid for zero-inflated variables. To accommodate zero-inflated and highly skewed variables, the truncated Gaussian copula models[10] have been introduced by incorporating an additional truncation mechanism, as follows:

Definition 1 (Truncated Latent Gaussian Copula Model)**.** *A random vector $\boldsymbol{y} \in \mathbb{R}^p$ satisfies the truncated latent Gaussian Copula model if there exists a random vector $\boldsymbol{y}^* \sim NPN(\mathbf{0}_p, \boldsymbol{\Sigma}, \boldsymbol{g})$ and constants D_j, $j = 1, \ldots, p$ such that $Y_j = I(Y_j^* > D_j)Y_j^*$ where $I(\cdot)$ is an indicator function. We then denote $\boldsymbol{y} \sim TLNPN(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{g}, \boldsymbol{D})$.*

The latent correlation matrix $\boldsymbol{\Sigma}$ of $TLNPN(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{g}, \boldsymbol{D})$ is estimated using Kendall's τ. The sample Kendall's τ between the jth and kth variables is defined as:

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \le i \le i' \le n} \operatorname{sign}(Y_{ij} - Y_{i'j}) \operatorname{sign}(Y_{ik} - Y_{i'k}).$$

There exists[10,11] an increasing bridge function G defined so $G(\Sigma_{jk}) = E(\hat{\tau}_{jk}) = \tau_{jk}$ where Σ_{jk} is an element of $\boldsymbol{\Sigma}$ corresponding to variables Y_j and Y_k. The bridge function G for two truncated variables is defined as:

$$G_{TT}(\Sigma_{jk}; \Delta_j, \Delta_k) = -2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \boldsymbol{\Sigma}_{4a}) + 2\Phi_4(-\Delta_j, -\Delta_k, 0, 0; \boldsymbol{\Sigma}_{4b}),$$

where $\Delta_j = f_j(D_j)$ and $\Phi_4(a_1, a_2, a_3, a_4; \boldsymbol{\Sigma}_4)$ denotes the CDF of 4-dimensional Gaussian with zero mean and correlation matrix $\boldsymbol{\Sigma}_4$ evaluated at $\boldsymbol{a} = (a_1, a_2, a_3, a_4)^\top$. The correlation matrices $\boldsymbol{\Sigma}_{4a}$ and $\boldsymbol{\Sigma}_{4b}$ are given by

$$\boldsymbol{\Sigma}_{4a} = \begin{pmatrix} 1 & 0 & 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} \\ 0 & 1 & \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} & 1 & -\Sigma_{jk} \\ -\Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & -\Sigma_{jk} & 1 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{4b} = \begin{pmatrix} 1 & \Sigma_{jk} & 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} \\ \Sigma_{jk} & 1 & \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} \\ \Sigma_{jk}/\sqrt{2} & 1/\sqrt{2} & 1 & \Sigma_{jk} \\ 1/\sqrt{2} & \Sigma_{jk}/\sqrt{2} & \Sigma_{jk} & 1 \end{pmatrix}.$$

Using the bridge function G, we can consistently[10,11] estimate the latent correlation matrix as $\hat{\Sigma}_{jk} = G^{-1}(\hat{\tau}_{jk}; \hat{\Delta}_j, \hat{\Delta}_k)$, where $\hat{\Delta}_j$ is the moment estimator as $\hat{\Delta}_j = \Phi(\hat{\pi}_j)$ and $\hat{\pi}_j = n^{-1} \sum_{i=1}^{n} 1(Y_{ij} = 0)$ is the sample proportion of zeros of the jth variable. Below, we see how the observed variable Y_j is modeled as a latent standard Gaussian variable, Z_j, truncated by Δ_j, where $\Phi(\Delta_j) = \pi_j$ is the probability of the jth variable taking zero, which is estimated by the sample proportion of zeros,

$$Y_j = I(Y_j^* > D_j)Y_j^* = I\{g(Y_j^*) > g(D_j)\}Y_j^* = I\{Z_j > \Delta_j\}Y_j^* = I\{\Phi(Z_j) > \Phi(\Delta_j)\}Y_j^* = I\{\Phi(Z_j) > \pi_j\}Y_j^*.$$

The latent Gaussian copula model for binary type data was first introduced in 2017 to model dependence among discrete Arabidopsis gene data.[11] The TLNPN model was introduced in 2020 along with the rank-based estimators for the latent correlation matrix, and it was found useful for modeling gene-expression and micro-RNA data.[10] The TLNPN model has shown to be useful when performing discriminant analysis for microbiome data due to its ability to model dependence among zero-inflated variables.[14] At the same time, zero-inflated and hurdle models are also popularly used to model zero-inflated data. However, a lack of research has been done comparing the TLNPN model to the other zero-inflated models and investigating the characteristics of data in which the TLNPN model performs better than the other models.

Definitions for Simulation Studies and Real Data Analyses

We examine the performance and robustness of ZINB, HNB, and TLNPN models using synthetic datasets across various conditions. We simulate data from each of the three populations—ZINB, HNB, and TLNPN—and, in settings two and three, evaluate performance by calculating the Wasserstein distance between test data independently generated from an assumed population model and data generated from the corresponding fitted model. The Wasserstein distance measures the distance between two probability distributions and is used for goodness-of-fit and statistical inference between two probability distributions where a lower distance implies a better fit.[15] Let μ and ν denote the probability measures corresponding to the distributions of the random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Also, let γ be a coupling, which is a probability measure defined on the product space of the probability spaces of $\boldsymbol{x}$ and $\boldsymbol{y}$, with marginals μ and ν. At the population level, the Wasserstein distance is defined as

$$W_p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left\{ \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma} d(\boldsymbol{x}, \boldsymbol{y})^p \right\}^{1/p},$$

where $\Gamma(\mu, \nu)$ is the set of all couplings. At the sample level, the Wasserstein distance between multivariate dataset $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ is defined as

$$W_p(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \inf_{\theta \in \mathcal{S}_n} \left(\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_{\theta(i)}\|^p \right)^{1/p},$$

where θ is a permutation in the symmetric group $\mathcal{S}_n$, the set of all n-permutations. Since $\mathcal{S}_n$ contains $n!$ permutations, we approximate the computation of Wasserstein distances in our numerical study using the network simplex algorithm,[16] implemented in the R package `transport`. As a summary measure of our results, we use arithmetic mean change (AMC). We use AMC because we want to measure the real relative difference between the performance of the models under varying scales of data. Furthermore, AMC is a symmetric measure of relative change, so a 10% improvement and 10% decline in performance from the HNB model to the TLNPN model are both captured by AMC, for example. Let $\omega_{\text{HNB}} = W_p(Y, \hat{Y}_{\text{HNB}})$ and $\omega_{\text{TLNPN}} = W_p(Y, \hat{Y}_{\text{TLNPN}})$ where $\hat{Y}_{\text{TLNPN}}$ is a simulated multivariate dataset generated from the TLNPN model, $\hat{Y}_{\text{HNB}}$ is a simulated multivariate dataset generated from the HNB model, and Y is a multivariate dataset generated from the true model. The AMC of the Wasserstein distance generated between the HNB model and the true model to the Wasserstein distance generated between the TLNPN model and the true model is given by:

$$\text{AMC} = \frac{\omega_{\text{TLNPN}} - \omega_{\text{HNB}}}{(\omega_{\text{TLNPN}} + \omega_{\text{HNB}})/2}. \qquad \textbf{Equation 3.}$$

Therefore, a positive AMC implies that the HNB model performs better, a negative AMC implies the TLNPN model performs better, and AMC=0 implies that the models perform the same.

We explore performance of ZINB, HNB, and TLNPN under three simulation settings. In setting one, we aim to validate previous findings on the robustness of univariate ZINB and HNB models to model misspecification, focusing specifically on varying levels of

zero inflation or deflation and differing proportions of sampling and structural zeros. In this simulation setting, we compare model performance using Akaike Information Criterion (AIC) since they are both discrete distributions and we seek to replicate previous studies.[1] AIC is a measure of model fit based on the likelihood function, with a penalty term that increases linearly with the number of model parameters k. It is defined as

$$\text{AIC} = 2k - 2\ln L,$$

where L is the likelihood function. Smaller AIC values suggest a more favorable model fit. It is a standard measure to compare two models; however, when comparing discrete and continuous models, it is biased towards the continuous model since likelihood values in continuous models are usually higher than probabilities in discrete models. Accordingly, we use AIC to compare the HNB and ZINB models but use Wasserstein distance to compare the HNB and TLNPN models. In simulation setting two, we compare the HNB model to the TLNPN model under HNB population data. We seek to evaluate whether the proportion of zeros or the dependence among the variables has an impact on the relative performance of each model. In simulation setting three, we again compare the HNB and TLNPN models except with TLNPN population data. We again seek to understand how zero-proportion and dependence among variables affects relative model performance.

In the multivariate settings (settings two and three), we apply the following autoregressive (AR) and geometrically decaying eigenvalues (GD) correlation structures to induce dependence between zero-inflated variables. The AR correlation structure is given by:

$$\boldsymbol{\Sigma} = [\rho^{|j-j'|}]_{1\le j,j'\le p}. \qquad \textbf{Equation 4.}$$

The covariance matrix of the GD structure is given by $\boldsymbol{\Sigma} = \boldsymbol{\Gamma N \Gamma}^T$ where $\boldsymbol{N} = \text{diag}\{\nu_j\}_{j=1}^{p}$ is a diagonal matrix with geometrically decaying eigenvalues defined by:

$$\nu_j = \frac{5(\rho^{j-1} - \rho^j)}{1-\rho^p}, \quad j = 1, \ldots, p, \qquad \textbf{Equation 5.}$$

where a lower value of ρ leads to higher correlations (in the absolute value sense) between the covariates, and $\boldsymbol{\Gamma}$ is uniformly generated from the orthogonal group[17] of order p, OG_p, where $OG_p = \{\boldsymbol{O} \mid \boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I}, \boldsymbol{O} \in \mathbb{R}^{p\times p}\}$ is the set of all $p \times p$ orthogonal matrices. The simulation settings are detailed as follows.

SETTING ONE: COMPARING THE UNIVARIATE ZINB AND HNB MODELS

In this setting, we seek to replicate previous findings comparing the univariate ZINB and HNB models under model misspecification and varying zero proportion.[1] We simulated $n = 500$ data points using covariate X_i where $X_i \sim N(0,1)$, $i = 1, \ldots, 500$, $\ln(\mu_i) = \beta_0 + \beta_1 x_i$, $\text{logit}(\pi_{Z,i}) = \gamma_0 + \gamma_1 x_i$ and $\text{logit}(\pi_{H,i}) = \gamma_0 + \gamma_1 x_i$ as in **Equation 1** and **Equation 2**. We performed simulations under three parameter conditions controlling for $\beta_0 = \ln(12)$, $\beta_1 = 2$, $\gamma_1 = 2$, and $r = 0.5$ at 20% zero-proportion, 40% zero-proportion, and 60% zero-proportion, which we adjusted using the γ_0 parameter under both the ZINB model and HNB model. We fix $\beta_1 = \gamma_1 = 2$ to replicate previous findings[1] and fix $\beta_0 = \ln(12)$ and $r = 0.5$ to ensure there is a significant difference in the proportion of structural and sampling zeros under the HNB and ZINB models. We then fit each model to the simulated data and compare the model fit through AIC.

To further investigate the impact that zero-deflation had on relative model performance, we conducted a follow-up simulation study. We simulated $n = 700$ data points using covariate $X_i \sim N(0,1)$, $i = 1, \ldots, 700$, $\ln(\mu_i) = \beta_0 + \beta_1 x_i$, and $\text{logit}(\pi_{H,i}) = \gamma_0 + \gamma_1 x_i$ where $\beta_0 = \ln(6/7)$, $\beta_1 = 0.1$, $\gamma_1 = 0$, and $r = 2$ in **Equation 2**. We varied $\pi_{H,i}$ from 0.08 to 0.7 by adjusting the γ_0 parameter. In this case, under the standard negative binomial distribution with $\mu = \ln(6/7)$ and $r = 2$, the probability of $Y = 0$ is 0.5. We fixed $\beta_0 = \ln(6/7)$ and $r = 2$ in order to make the probability of $Y = 0$ under the standard negative binomial distribution equal to 0.5, so there was a large range of zero-proportions that would be considered zero-deflation. Furthermore, we fixed $\beta_1 = 0.1$ and $\gamma_1 = 0$ because we were mainly interested in the effect of zero-deflation, so we did not want the covariate to have a large impact on the mean or probability of a structural zero. For each iteration, corresponding to a different proportion of zeros, we fit the data using both HNB and ZINB models and compared their AIC values.

We confirmed previous findings and found that when the proportion of sampling and structural zeros differed significantly and β_1 and γ_1 had high values, then the models were sensitive to model misspecification[1] as seen in **Figure 3**. We see that under each propor-

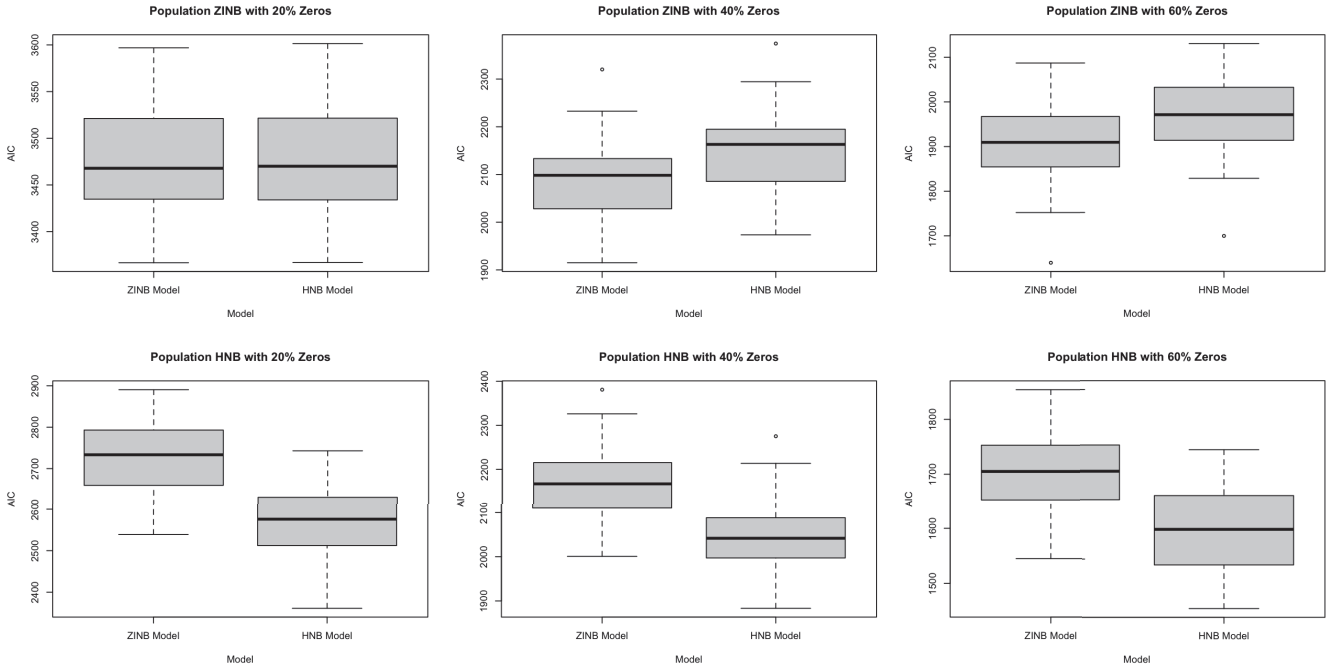

Figure 3. The box plots show how the AIC of the ZINB and HNB models compare with one another under varying conditions. On the top row, the population data was generated from the ZINB model, and on the bottom row, the population data was generated from the HNB model. In this case, $\beta_1 = \gamma_1 = 2$. Each column of box plots corresponds to a different proportion of zeros: 20%, 40%, and 60%. We see in the top row that the ZINB model outperforms the HNB model with the exception of the setting with 20% zero-proportion, and on the bottom row, the HNB model outperforms the ZINB model. We see this result because given a x_i with a large negative magnitude, the ZINB model will predict a sampling zero where the HNB model would not. For example, given $x_i = -5$, $\pi_Z + (1 - \pi_Z)(\frac{r}{\mu_i + r})^r \approx 1$ in **Equation 1**; however, $\pi_H \approx 0$ in **Equation 2**.

tion of zeros (20%, 40%, and 60%), when $\beta_1 = \gamma_1 = 2$, the true model far outperformed the other in terms of AIC (with the exception of when the population data was generated by the ZINB model and the zero-proportion was 20%). In a second simulation study comparing the HNB and ZINB models, our results show that under conditions of zero deflation—where the probability of $Y = 0$ is less than 0.5—the HNB model significantly outperforms the ZINB model in AIC. The difference in AIC values grows exponentially as the proportion of zeros falls below the probability of $Y = 0$ under the standard negative binomial distribution, as illustrated in **Figure 4**. However, it seems that the ZINB model is able to account for moderate zero-deflation without an impact on model performance.

In this setting, we found that when $\beta_1 = \gamma_1 = 2$, the ZINB and HNB were vulnerable to model misspecification as seen in **Figure 3**. We conclude that this trend is due to the fact that the ZINB model will predict a sampling zero given x_i with a large, negative magnitude whereas the HNB model would never predict a sampling zero, causing the difference in AIC. Furthermore, we also observed that in cases of zero-deflation, the HNB model far outperformed the ZINB model, which is displayed in **Figure 4**. However, the ZINB model seemed robust to moderate levels of zero-deflation. This robustness of the ZINB model to moderate zero deflation arises from its ability to adjust parameters such as the dispersion parameter, compensating for a lower-than-expected proportion of zeros. However, as the zero-deflation intensifies, the difference in AIC begins to grow since the ZINB model cannot predict zeros at a probability below that of a standard negative binomial distribution.

SETTING TWO: COMPARING THE HNB AND TLNPN MODELS (WITH HNB POPULATION DATA)

This setting aims to empirically compare the goodness-of-fit of the HNB and TLNPN models when the population data are generated from the HNB model. We seek to investigate how different parameters of the HNB population model affect the relative performance of the TLNPN model to discover both its strengths and weaknesses. We do not consider the ZINB model because it cannot model zero-deflation, which is one of the conditions that we investigate. We also consider the HNB model fitted with and without covariates; however, in most biomedical datasets, covariates are unavailable. We set $n = 1200$ and $p = 5$ and generate covariates $\boldsymbol{x}_i$,

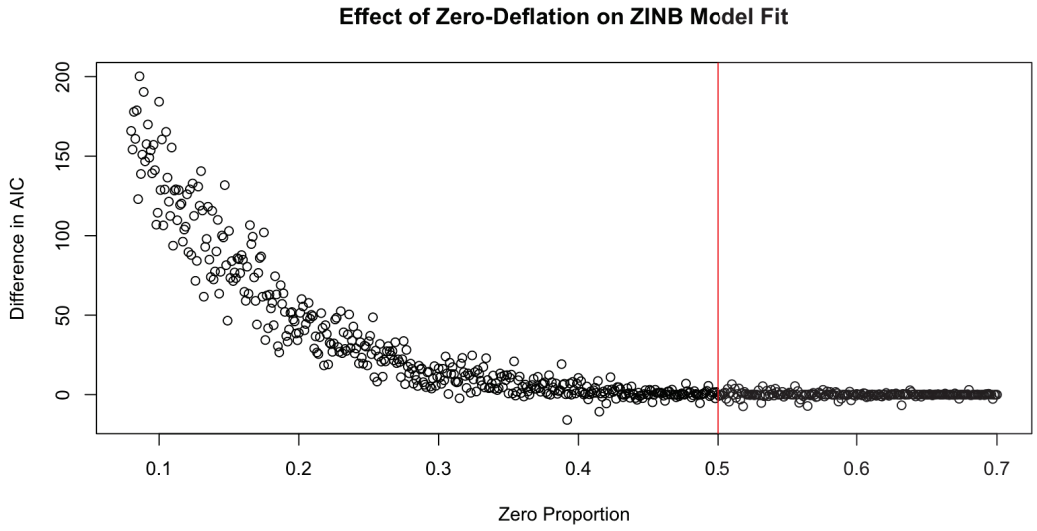

Figure 4. In cases of extreme zero-deflation, the HNB far outperformed the ZINB model. We simulated HNB population data under varying levels of zero-deflation and inflation and fitted both the HNB and ZINB model. We found the difference in AIC between the two models corresponding to each proportion of zeros. The y-axis displays the ZINB model AIC minus the HNB model AIC. Under the standard negative binomial model with $\mu = \ln(6/7)$ and $r = 2$, the probability that $Y = 0$ is 0.5 (the red line). We see that as the proportion of zeros declines below 0.5, the difference in AIC grows at an increasing rate.

$i = 1, \ldots, n$, independently and identically from the multivariate Gaussian with zero mean and covariance matrix given in **Equation 4** and **Equation 5**. We then set $\ln(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$ and $\text{logit}(\pi_{Hij}) = \gamma_0 + \gamma_1 x_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, p$ as in **Equation 2**. We set $\beta_0 = 2.75$ and $r = 6$ for all the simulations, which allows us to evaluate the impact of zero-inflation and zero-deflation. The parameter β_1 controls the impact that the covariate has on μ_{ij} and thus, affects the scale and dependence among the variables. It was set at 0, 1, and 2 in order to evaluate the impact of x_{ij} having no effect on μ_{ij} to x_{ij} having a large effect on μ_{ij}. The parameter γ_0 controls the zero-proportion of the variable and was set at $\ln(1/20)$, $\ln(1/9)$, and $\ln(1/3)$ since we are interested in whether zero-inflation or zero-deflation impacts relative model performance. We also varied the parameter γ_1, which controls how the covariate affects the probability of a structural zero, which impacts the variance of the data. This parameter was set at -0.8, 0.8, and 0. These values were chosen to evaluate the effect of x_{ij} having a negative, positive, and no relationship with $P(Y_{ij} = 0)$, respectively. Finally, we also investigated the impact of the parameter ρ, which controls the amount of correlation between the covariates under both the AR and GD correlation structures, and thus, it affects the dependence among the HNB variables. The parameter ρ was set at 0.01, 0.3, 0.7, and 0.9. We wanted to repeat this simulation study with a GD correlation structure because it can create negative correlation between covariates, unlike the AR correlation structure, and it more closely resembles correlation matrices found in real life.

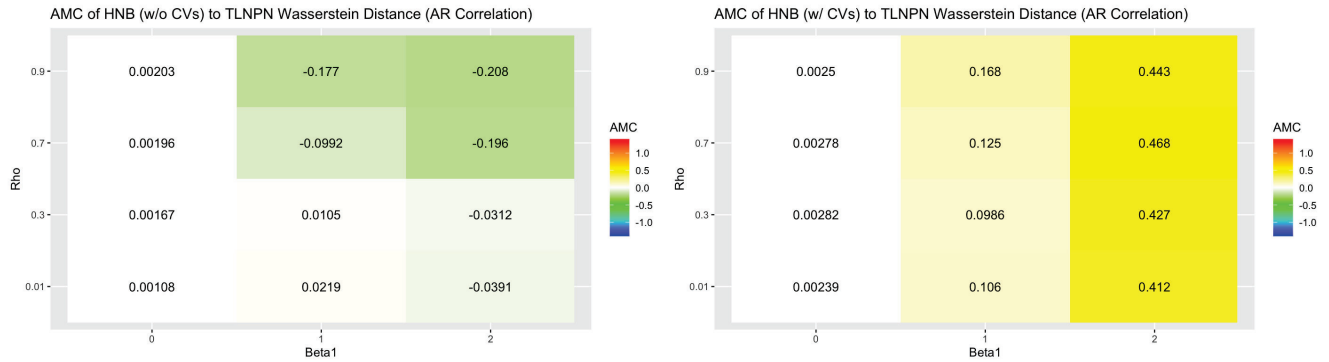

Figure 5. These heatmaps compare the performance of the HNB model to the TLNPN model through our summary measure arithmetic mean change (AMC) given in **Equation 3**. Positive AMC, represented by warm colors, indicates the HNB model performing better, and negative AMC, represented by cool colors, indicates the TLNPN model performing better. When the HNB model is fitted without covariates, we see that as ρ and β_1 increase, the TLNPN model outperforms the HNB model. However, when the HNB model is fitted with covariates, the HNB model outperforms the TLNPN model most as ρ and β_1 increase.

As a goodness-of-fit measure, we consider 5-fold cross-validated prediction error. In particular, we randomly split $n = 1200$ observations into five equal-sized folds, using four of the folds for training and keeping the remaining one for testing. The HNB and TLNPN models are fitted to the training data, and from each fitted model, we simulate a dataset of 240 observations for prediction and obtain the Wasserstein distance between the simulated data from the fitted model and the test data. The described process is repeated for each fold, and we average the resulting five Wasserstein distances to obtain the cross-validated prediction error, which will be used as our performance measure. We repeat this procedure for 30 replicated datasets and summarize the results in **Figure 5** through **Figure 12**.

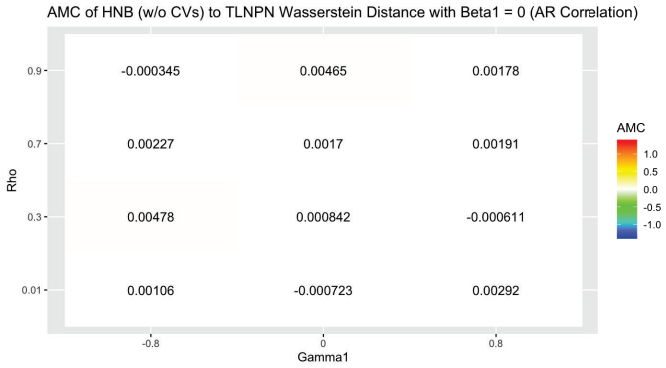

Figure 6. When $\beta_1 = 0$, the HNB model and TLNPN model perform the same. We see this result because when $\beta_1 = 0$, then the HNB variables are practically independent (the γ_1 parameter could still incur a minor amount of dependence between the variables). Therefore, the TLNPN model does not have an advantage in fitting the multivariate distribution as there is very little dependence among the variables for it to model.

We first investigate the interaction between β_1 and ρ and its impact on relative model performance. When we account for whether covariates were included in the fitting of the HNB model, we find starkly different results as shown in **Figure 5**. We see that when covariates aren't considered in the fitting of the HNB model, we see results consistent with our predictions that a high ρ and high β_1 parameter would improve the relative performance of the TLNPN model. We also see that when $\beta_1 = 0$ or $\rho = 0.01$, the models perform nearly the same. However, when covariates are considered in fitting the HNB model, we see notably different results. Still, the TLNPN and HNB models perform nearly the same when $\beta_1 = 0$ with the AMC of the Wasserstein distances being approximately 0 regardless of ρ. However, as β_1 increases, the relative performance of the TLNPN model against the HNB model worsens. Furthermore, it seems that as ρ increases, the TLNPN model performs worse against the HNB model fitted with covariates. We conducted a follow-up simulation study to investigate this trend.

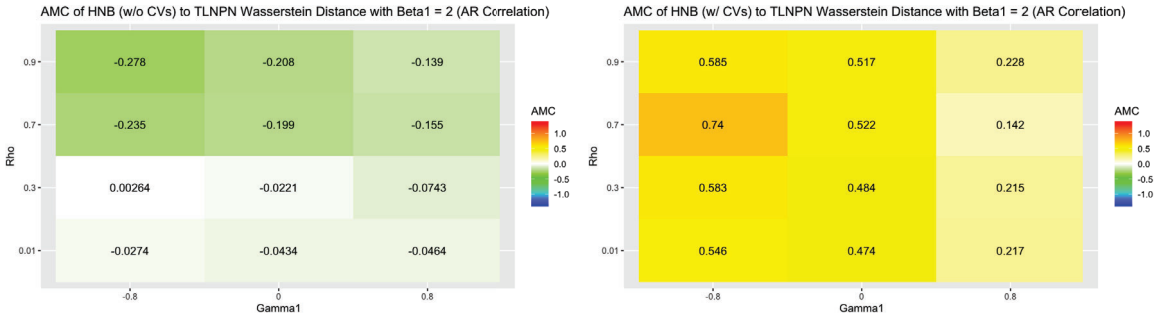

Figure 7. The heatmaps show the relative performance of the TLNPN model to the HNB model under different values of ρ and γ_1. We see that when covariates are not considered when fitting the HNB model, then the TLNPN model performs best relative to the HNB model when $\gamma_1 = -0.8$ and $\rho = 0.9$ (left). When $\gamma_1 = -0.8$, as the covariate, x_{ij}, increases, the mean of the model increases while the probability of a structural zero decreases. Since the covariates for each variable are correlated, this results in stronger correlation among the zero-inflated variables. However, when $\gamma_1 = 0.8$ an observation is more likely to contain a variable with a high value and one that is equal to 0, reducing the dependence among the variables leading to an increase in the AMC. We also see that when covariates are considered in fitting the HNB model, then the TLNPN model performs worse relative to the HNB model as γ_1 decreases and generally as ρ increases (right).

The follow-up simulation study followed the same process as the first except we also measured the marginal Wasserstein distance between the test data and the simulated data and the correlation matrix of the test data and the correlation matrix of the simulated data. In this study, we considered the HNB model fitted with covariates. We found that as β_1 increases from 1 to 2 (when $\rho = 0.9$), the

TLNPN model will further underestimate the correlation between the zero-inflated variables as shown in **Figure 8**. We also found that the one-dimensional Wasserstein distances of the marginal TLNPN data to the marginal test data were greater than that of those generated by HNB data as displayed in **Figure 9**.

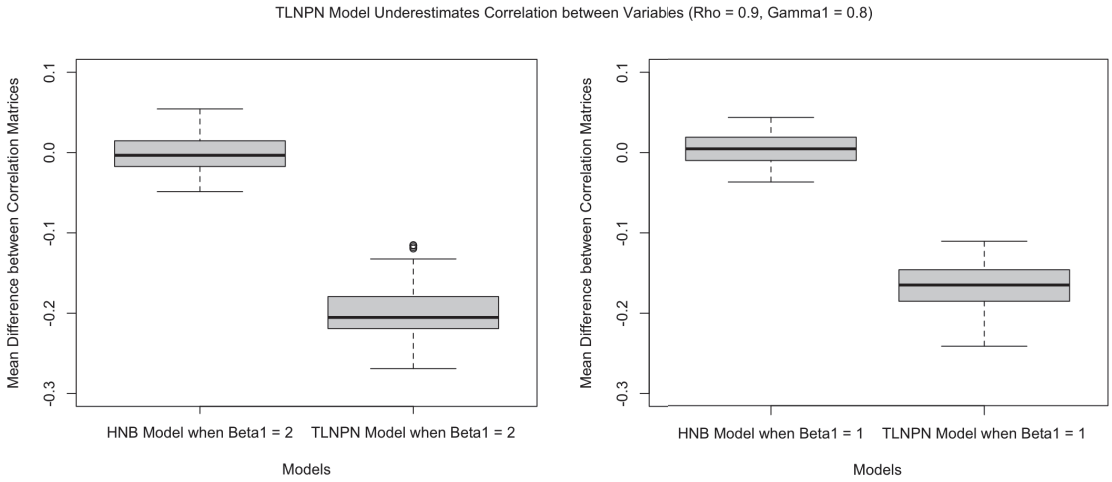

Figure 8. In these box plots, the y-axis measures the mean difference between the values of the correlation matrices between the data produced by the model and the test data. In this figure, the HNB model was fitted with covariates. These box plots show that as β_1 increases, the TLNPN model increasingly underestimates correlation between variables. We suspect this occurs because the latent Gaussian variables in the TLNPN model must fit a latent correlation matrix such that the joint distribution will have data points with variables equal to large positive values and variables equal to 0 on account of the zero-inflation.

We found that the zero-proportion of the variables, controlled by γ_0, did not have an impact on relative model performance between the HNB and TLNPN models. Furthermore, we also investigated the interaction between ρ and γ_1 with regard to the relative model fit between the TLNPN and HNB models when controlling for β_1. We see in **Figure 6** and **Figure 23** that regardless of whether covariates are included in the HNB model fitting, when $\beta_1 = 0$, γ_1 has no impact on the relative performance of the TLNPN model. When $\beta_1 = 2$, we see a trend in **Figure 7**. When covariates are not considered, the TLNPN model performs best when $\gamma_1 = -0.8$ and $\rho = 0.9$. Additionally, as ρ increases, the performance of the TLNPN model against the HNB model improves. When covariates are considered, the trend is reversed where the TLNPN performs worse when $\gamma_1 = -0.8$ and ρ is high. We conducted a follow-up simulation study to investigate this trend.

From our previous follow-up simulation study, we see that as γ_1 increases, the one-dimensional Wasserstein distance between the TLNPN data and the test data decreases as seen in **Figure 9**. To explain the differences in the one-dimensional Wasserstein distances, we conducted another follow-up simulation study in which we compared the distributions of the test data to the HNB and TLNPN simulated data when $\gamma_1 = -2$ and when $\gamma_1 = 2$, and the results are displayed in **Figure 10**. We see that when $\gamma_1 = -2$, we find much higher residuals between the TLNPN and test data; however, when $\gamma_1 = 2$, these residuals decrease dramatically.

The results of this simulation study presented thus far have been generated from covariates following an AR correlation structure; we found similar results from covariates generated from the GD correlation structure. Again, note that in a GD correlation structure, as ρ decreases, the correlation (in an absolute sense) between the covariates tends to increase.

We see in **Figure 11**, that again, there is an interaction between ρ and β_1 such that when ρ is small and β_1 is large, the TLNPN model outperforms the HNB model fitted without covariates. We again see similar results when the HNB model is fitted with covariates where the HNB model outperforms the TLNPN model most when $\beta_1 = 2$ and $\rho = 0.01$. For both scenarios, we see that when $\beta_1 = 0$, the models perform nearly identically.

As with the AR correlation structure, we again see in **Figure 12** an interaction between β_1 and γ_1 where the effect of γ_1 only becomes clear when $\beta_1 \neq 0$ since the models perform nearly identically when $\beta_1 = 0$ regardless of ρ or γ_1. When $\beta_1 = 2$, we begin to see a familiar pattern. When covariates are not considered when fitting the HNB model, the TLNPN model outperforms the HNB model

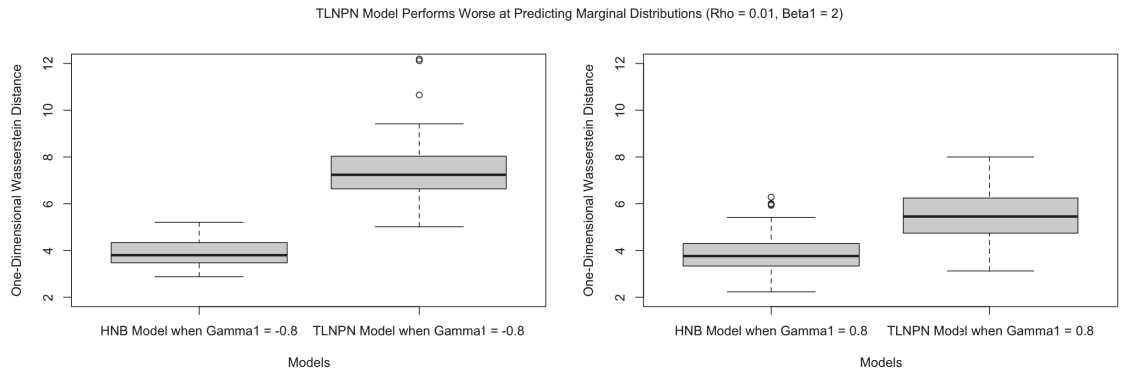

Figure 9. These box plots show the impact that γ_1 has on the marginal performance of the TLNPN model. The y-axis displays the one-dimensional Wasserstein distance between the first variable of the test data and the first variable of the simulated data. We see that as γ_1 increases, the Wasserstein distance between the TLNPN data and test data decreases. This trend is a result of higher, more extreme values become less probable as γ_1 increases, which the TLNPN model struggles to predict since the variance of the HNB model increases as μ_{ij} increases. The HNB model was fitted with covariates, and we see that the marginal Wasserstein distance between the HNB simulated data and the test data stays approximately the same as γ_1 increases.

most when $\rho = 0.01$ and $\gamma_1 = -0.8$ where as ρ and γ_1 decrease, the better the TLNPN model performs relative to the HNB model. However, when covariates are considered when fitting the HNB model, the relative performance of the TLNPN model declines as ρ and γ_1 decrease.

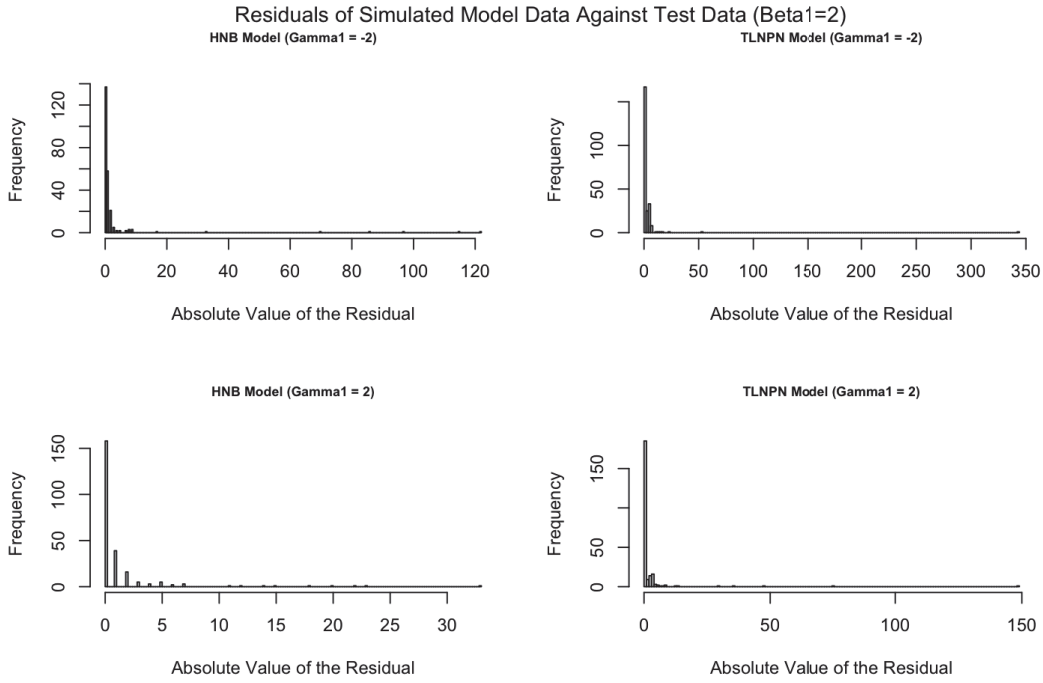

Figure 10. These histograms display how γ_1 affects the marginal performance of the HNB and TLNPN models in the first and second column respectively. The first row shows the difference between the sorted values of the simulated and test data when $\gamma_1 = -2$, and the second row shows the case when $\gamma_1 = 2$. As γ_1 increases, the graphs show the residuals decreasing, particularly for the TLNPN data, which occurs because when γ_1 increases, given $\beta_1 > 0$, extreme values become less likely because as the covariate increases, both μ_{ij} and $\pi_{H,ij}$ increase.

The first main result of our second simulation setting is displayed in **Figure 5** where we see that when the HNB model is fitted without covariates, the TLNPN model performs best when $\beta_1 = 2$ and $\rho = 0.9$ under the AR correlation structure. We attribute this to the ability of the TLNPN model to account for the correlation between the latent Gaussian variables, which can be influenced through the correlation between the covariates. However, for the correlation between the covariates to have an impact on the latent correlation, the β_1 parameter, which controls the impact the covariates have on the mean, has to be nonzero. We see that regardless of

ρ, when $\beta_1 = 0$, the models perform nearly identically because the impact of the correlation between the covariates has no bearing on the correlation of the latent variable since the covariates have no effect on the mean. Similarly, when $\beta_1 = 2$ and $\rho = 0.01$, we see that the models perform nearly the same, because there is a lack of dependence among the covariates and thus, the variables. Therefore, under the AR correlation structure, when both ρ and β_1 are large, we see the best relative performance of the TLNPN model because the variables are more highly dependent on each other since the covariates are highly correlated and the covariates have a large impact on μ.

In **Figure 5**, we also observe that when the HNB model is fitted with covariates, it outperforms the TLNPN model most when both β_1 and ρ are high. We conducted a follow-up simulation study to investigate this trend; the results of which are summarized in **Figure 8** and **Figure 9**. We see in these figures that the TLNPN model underestimates the correlation among the HNB variables and performs worse than the HNB model fitted with covariates on the marginal level. We attribute this trend to the TLNPN model estimating the latent correlation between the latent Gaussian variables through a formula that utilizes Kendall's τ. In contrast, the HNB model was able to much more accurately capture the correlation between the variables. We attribute this to the HNB model using the covariates of the test data when simulating data from the fitted model, resulting in more accurate predictions than those of the TLNPN model, particularly for extremely high values. The TLNPN model can only predict values within its training dataset, which makes it vulnerable to modeling datasets with extreme, outlier values, which are much more probable when $\beta_1 = 2$ as compared to 1 or 0.

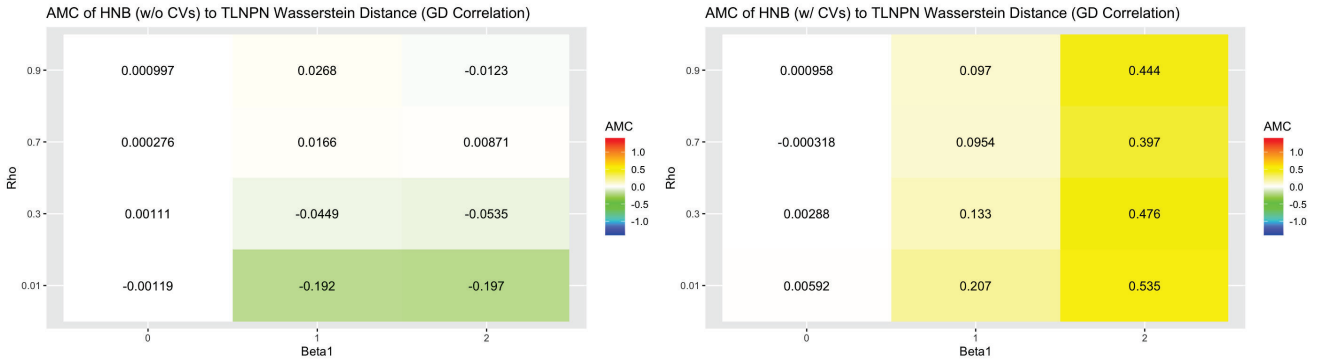

Figure 11. We see in the first heat map that the TLNPN model outperforms the HNB model (fitted without covariates) most when $\rho = 0.01$ and $\beta_1 = 2$. This occurs because under the GD correlation structure, when $\beta_1 = 2$ and $\rho = 0.01$, there is a stronger dependence among the zero-inflated variables. In the second heat map, the HNB model is fitted with covariates. We see that the TLNPN model performs worst relative to the HNB model when $\beta_1 = 2$ and $\rho = 0.01$ since the TLNPN model underestimates the correlation among the variables and performs worse at modeling the marginal distributions.

In **Figure 6**, we see that when $\beta_1 = 0$, the γ_1 parameter seems to have no effect on relative model performance. This is a result of the lack of dependence and extreme values among the variables that results when $\beta_1 = 0$, so the parameter γ_1 can only have a limited impact on the dependence and variance of the variables. However, when $\beta_1 = 2$, we see a clear pattern emerge both when the HNB model is fitted with covariates and when it is not as shown in **Figure 7**. When covariates are not considered when fitting the HNB model, the TLNPN model outperforms the HNB model most when $\rho = 0.9$ and $\gamma_1 = -0.8$. We attribute this trend to the fact that β_1 is always non-negative in our simulations, therefore, if an increase in the covariate both increases the mean and decreases the probability of a zero, then the resulting latent correlation, calculated from Kendall's τ will be much stronger, which the TLNPN model accommodates.

Despite this, when covariates are considered when fitting the HNB model, the pattern reverses, and the TLNPN model performs worse when $\gamma_1 = -0.8$ and ρ is large. We conducted a follow-up simulation study to investigate, and the result is displayed in **Figure 10** where we see that as γ_1 increases, the residuals between the simulated data and the test data greatly reduces, particularly for the TLNPN data. Therefore, there are two trends at work that cause the TLNPN model to perform worse against the HNB model fitted with covariates when $\gamma_1 = -0.8$ as compared to when $\gamma_1 = 0.8$ (given ρ is high and $\beta_1 > 0$). One, when $\gamma_1 = -0.8$, the probability of a structural zero decreases as the covariate, and therefore the mean of the distribution, increases. This increases the correlation between the zero-inflated variables, and the HNB model more accurately describes the correlation structure between the zero-inflated variables than the TLNPN model, which tends to underestimate the correlation between the variables. Two, as the mean of the HNB

distribution increases, the variance increases as well, which will increase the Wasserstein distance between test and simulated data. However, the HNB model is better equipped to predict higher values compared to the TLNPN model because that model uses the same covariates as the test data. We see that when $\gamma_1 = 0.8$, there is an improvement in the relative performance of the TLNPN model against the HNB model fitted with covariates modeling the marginal distribution compared to when $\gamma_1 = -0.8$ as seen in **Figure 9**. Therefore, the TLNPN model performs relatively better when $\gamma_1 = 0.8$ because it makes the occurrence of an extremely high data point less probable, which we see in **Figure 10** where as γ_1 increases, the size of the residuals between the marginal distributions decreases dramatically. We also found that the level of zero-inflation or deflation had no effect on the relative model performance. We conclude that this results from the ability of both models to account for both zero-inflation and zero-deflation.

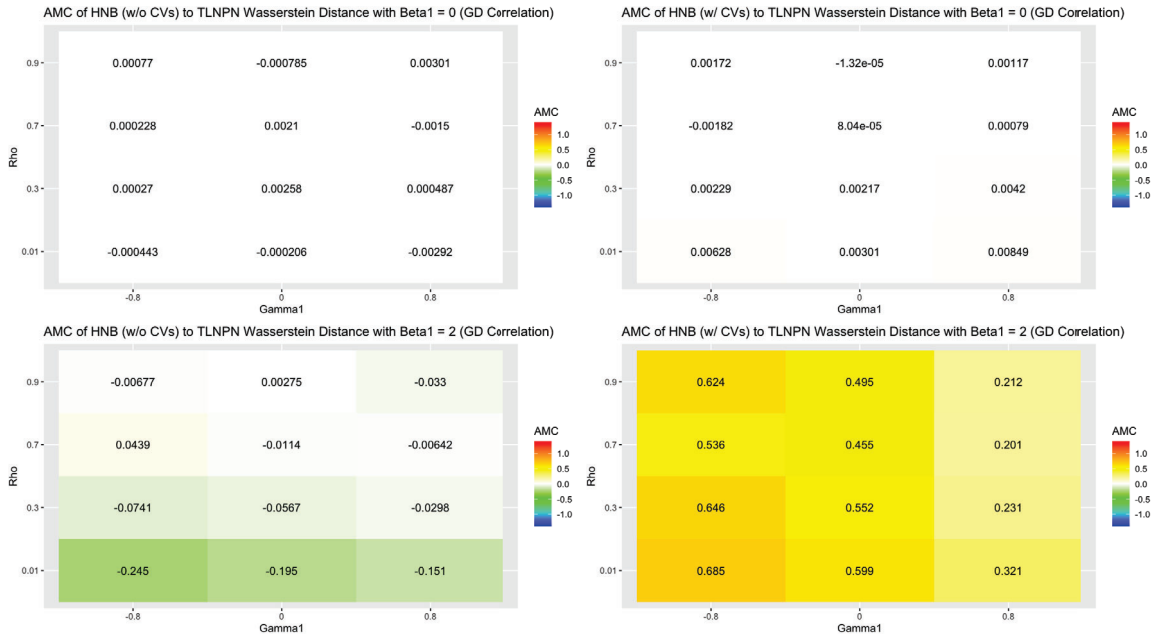

Figure 12. In the left column, the HNB model is fitted without covariates, and in the right column, the HNB model is fitted with covariates. On the top row, we consider when $\beta_1 = 0$, and on the bottom row, we consider when $\beta_1 = 2$. We see on the top row, that the TLNPN model and HNB model perform nearly identically when $\beta_1 = 0$ as the variables are practically independent and have much less variance compared to when $\beta_1 = 2$. We see on the lower left heat map that the TLNPN model outperforms the HNB model fitted without covariates most when $\rho = 0.01$ and $\gamma_1 = -0.8$ as a result of the increased dependence among the variables. We see in the lower right heat map, the TLNPN model performs worst relative to the HNB model when $\rho = 0.01$ and $\gamma_1 = -0.8$ since the TLNPN model underestimates the correlation among the variables and performs worse at modeling the marginal distributions.

We performed the simulation study where the covariates were generated from a GD correlation structure. Under the GD correlation structure, we see results investigating the interaction between ρ and β_1 in **Figure 11** where when the HNB model is fitted without covariates, the TLNPN model outperforms the HNB model when $\rho = 0.01$ and $\beta_1 = 2$. We conclude that this results from the dependence among the HNB variables that results when the covariates are highly correlated and have a large impact on μ_{ij}, which the TLNPN model accounts for, but the HNB does not. We see that when the HNB model is fitted with covariates, the HNB model outperforms the TLNPN model most when $\beta_1 = 2$ and $\rho = 0.01$ since it more accurately describes the dependence structure and can better predict large values. For both scenarios, we see that when $\beta_1 = 0$, the models perform nearly identically as the zero-inflated variables have almost no dependence among each other, and both models fit the marginal distributions similarly well.

We also investigated the interaction between ρ and γ_1 under the GD correlation structure, which is presented in **Figure 12**. The interpretation of these results is the same as the interpretation of the results when the covariates are generated from an AR correlation structure; when $\beta_1 = 0$, then γ_1 has very little impact on the dependence structure of the zero-inflated variables, so it doesn't have an impact on the relative performance between the TLNPN and the HNB models regardless of whether the HNB model was fitted with covariates. However, when $\beta_1 > 0$ and $\gamma_1 < 0$, then the correlation between the zero-inflated variables will strengthen since the higher the mean of the distribution is, the lower the probability of a structural zero. When the HNB model is fitted without covariates, the TLNPN model outperforms the HNB model the most when $\rho = 0.01$ and $\gamma_1 = -0.8$; however, when the HNB model is fitted with

covariates, the opposite is true due to the HNB model better describing the dependence structure and marginal distributions of the zero-inflated variables.

SETTING THREE: COMPARING THE HNB AND TLNPN MODELS (WITH TLNPN POPULATION DATA)

In this simulation study, we again compare the performance of the TLNPN model to the HNB model, but use the TLNPN model as the true model using Quantitative Microbiome Profiling (QMP) data.[2] We evaluated the performance of both models under different conditions. The motivation behind this simulation study is to evaluate how the two models compare under varying parameters of the TLNPN population model, which include zero-proportion, latent correlation, and variance of the training data. We consider both AR and GD correlation structures for the latent correlation matrix. We consider the GD correlation structure in order to simulate correlation matrices that are commonly found in real-world biomedical datasets, and we consider the AR correlation structure in order to more directly evaluate the impact of ρ on the relative performance of the models. We used the empirical CDF of both the original (untransformed) data and the square root transformation of the data as the training data of the population TLNPN model since the QMP dataset is extreme scale data, and we seek to investigate whether the scale and skewness of the data impacts relative model performance. We also vary the proportion of zeros in the TLNPN data (ZP) to evaluate whether zero-proportion has an effect on relative model performance. Finally, we set the correlation parameter, ρ, at different values to evaluate if the dependence of the variables has an impact on relative model performance. For this study, we generated the Gaussian-level variables with a correlation matrix of $\mathbf{\Sigma}$, so $\mathbf{x}_i = (X_{i1}, X_{i2}, \ldots, X_{i5})^\top \sim N_p(\mathbf{o}, \mathbf{\Sigma}), i = 1, \ldots, n$ where $n = 1200$. Let $\hat{F}_j$ be the empirical CDF of the jth variable of the QMP data. We generate data such that $y_{ij} = \hat{F}_j^{-1} \circ \Phi(x_{ij}), i = 1, \ldots, n$ and $j = 1, \ldots, p$ where $p = 5$. In this study, we control for the amount of zeros by subsetting on five variables in the QMP dataset that had the desired zero-proportions, then selecting those to generate the TLNPN data. In particular, for the jth variable, $\hat{\pi}_j = \Phi(\Delta_j)$ where $D_j = g_j^{-1}(\Delta_j)$ and $g_j^{-1} = F_j^{-1} \circ \Phi$, which is how each D_j is selected this in simulation study. We then split the population data into five folds, which we used for five-fold cross validation. We fitted the TLNPN and HNB models to the training data and generated data from each model, and find the respective Wasserstein distances between the simulated data and the test data. The average between the five-folds is then found and recorded.

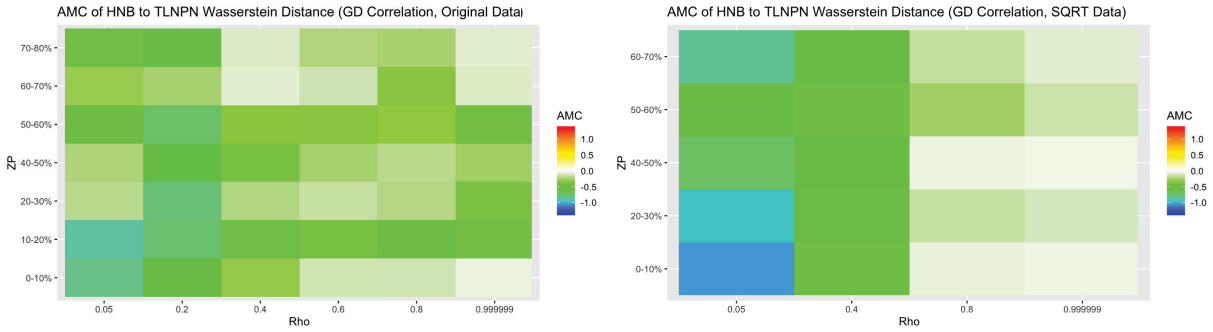

Figure 13. In these heat maps, we compare the performance of the TLNPN and HNB models under TLNPN population data. On the left, the TLNPN population model was trained on the untransformed QMP data. On the right, the TLNPN population model was trained on the square root of the QMP data. We see that in both cases, the TLNPN universally performs better regardless of the zero-proportion of the data or ρ. Although there is no clear pattern in the first heat map, we see in the second heat map, a clear pattern emerge: as ρ decreases and as zero-proportion decreases (when $\rho = 0.05$), the TLNPN model improves its performance relative to the HNB model.

In this simulation study, we investigate the impact that the zero-proportion and correlation of the TLNPN variables have on relative model performance. Our results for the GD correlation structure are presented in **Figure 13**. We see that when evaluating the models under the TLNPN variables distributed as the original, untransformed data, there is no clear pattern; the TLNPN model outperforms the HNB model in all cases, but it's not clear how zero-proportion or ρ impacts the AMC of the HNB Wasserstein distance to the TLNPN Wasserstein distance. However, when we use the square root of the data, we find a much clearer pattern. The lower ρ is when using the GD correlation structure, the better the TLNPN model performs relative to the HNB model.

In **Figure 14**, we present our results for the AR correlation structure. Here, our results are quite similar: the higher ρ is (meaning the higher the correlation between the latent Gaussian variables is), then the better the TLNPN model performs relative to the HNB model. Again, we see that zero-proportion does not reliably impact relative model fit except when $\rho = 0.999999$.

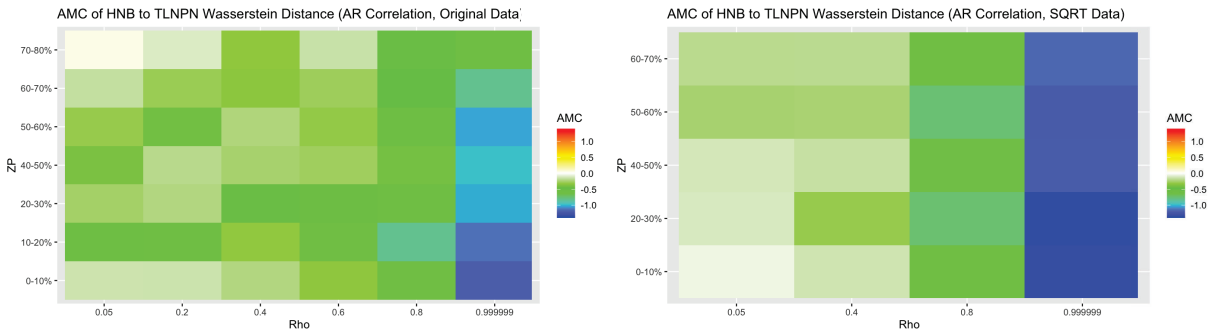

Figure 14. In these heat maps, we compare the performance of the TLNPN model to the HNB model where the population data was generated from the TLNPN model and the latent correlation matrix follows an AR correlation structure. On the left, the population TLNPN data was trained on the untransformed data, and on the right, the population TLNPN data was trained on the square root of the QMP data. We see in both heat maps, the TLNPN model universally outperforms the HNB model, and the TLNPN model performs best when $\rho = 0.999999$ (i.e., when there is strong dependence among the variables). Additionally, we see that the zero-proportion of the variables does not have a reliable effect on relative model performance except when $\rho = 0.999999$ where as zero-proportion decreases, the AMC decreases.

In **Figure 13**, we display results from the third simulation setting where we investigate the effect of zero-proportion, ρ (under the GD correlation structure), and the square-root data transformation on relative model performance. We see that in the square-root transformation, the TLNPN model outperforms the HNB model most when $\rho = 0.05$. We can attribute this to the higher correlation between the zero-inflated variables, which the TLNPN model is able to account for as opposed to the HNB model. However, it still seems that zero-proportion does not have an effect on the relative model performance. We attribute this to the fact that both models can handle zero-inflation or deflation. We display our results for the AR correlation structure in **Figure 14**, and see in both the untransformed CDF and square-root transformation CDF, the TLNPN model improves its performance against the HNB model as ρ increases. We again conclude that this is due to the TLNPN model's ability to model dependence among the zero-inflated variables.

REAL DATA ANALYSES

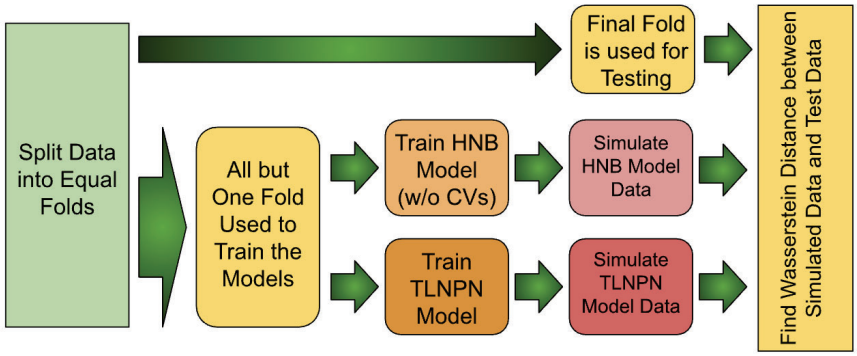

Figure 15. This figure displays a schematic illustration of real data analysis procedure.

In this section, we compare the Hurdle model and the truncated latent Gaussian copula model in their ability to describe real data examples. Our real data studies used datasets from a gut bacteria article[2] and gene-sequencing data (*https://www.10xgenomics.com*). We used regular validation, three-fold for the Quantitative Microbiome Profiling Data and five-fold for the gene sequencing data. We trained the HNB and TLNPN models on all but one fold, then simulated data from those models and found the Wasserstein distance between the simulated data from each model with the final fold. We considered 50 random splits, and our analysis process is graphically summarized in **Figure 15**. We summarize the relative performance of the models using AMC as given by **Equation 3**.

Quantitative Microbiome Profiling Data

As an example of real world zero-inflated data, we use Quantitative Microbiome Profiling (QMP) data.[2] This data measures the number of 101 different genera of gut bacteria in 135 people (29 with Crohn's Disease and 106 controls). We found that a limitation of fitting the HNB model was that the most popular R function used to fit the HNB model (from the `pscl` package) was limited to integers less than or equal to $2^{31} - 1$.[18] We had to rescale the data by taking the power of 0.851 of each data point and then rounding, which

makes the maximum value of the data $2^{31} - 1$. The first, second, third, and fourth quartiles of the zero-proportion of the 101 variables are 3.7%, 28.9%, 57.8%, and 79.3% respectively, and the data displays a high amount of skewness. The result of the analysis is displayed in **Figure 16**.

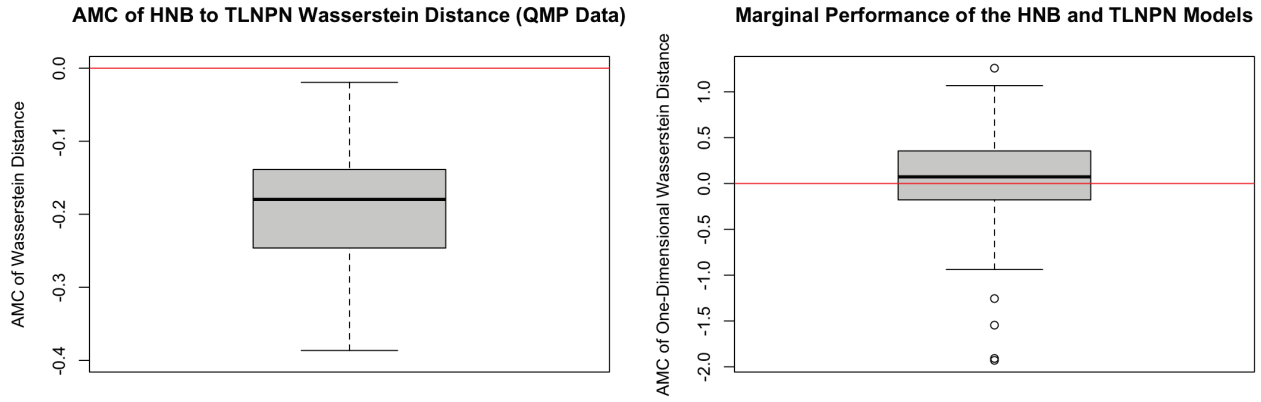

Figure 16. The left box plot shows the Arithmetic Mean Changes (AMC), defined in **Equation 3** of the HNB Wasserstein distance to the TLNPN Wasserstein distance, based on 50 random splits of QMP data. The red line at zero marks the reference: points above indicate a better HNB fit, and points below indicate a better TLNPN fit. The left panel shows that the TLNPN data had a lower 101-dimensional Wasserstein distance with the test data than the HNB data. We found the one-dimensional Wasserstein distance between each of the variables of the test data and each of the variables of the simulated data from the models. The right panel shows the AMC of the HNB to the TLNPN one-dimensional Wasserstein distance for each variable. Both models performed similarly on marginal distributions, but TLNPN model consistently outperformed HNB model on the joint distribution.

We found that the TLNPN model outperformed the HNB model in terms of p-dimensional Wasserstein distance in every replication. We also see in **Figure 16** a box plot of the AMC of the one-dimensional Wasserstein distances of the each of the 101 variables from the HNB model to the TLNPN model.

Our first real data analysis compared the performance of the TLNPN model against the HNB model using the QMP dataset; the results are displayed in **Figure 16** where we see that the TLNPN model outperformed the HNB model with regard to p-dimensional Wasserstein distance under every random split but performed similarly, on average, on a marginal level. We conclude that the difference between the Wasserstein distances was a result of the TLNPN model accounting for dependence among the variables whereas the HNB model does not model dependence among the variables. We see that the two models perform, on average, about the same when modeling the marginal distributions, which rules out the explanation that the TLNPN model outperformed the HNB model due to marginal fit.

To further emphasize this difference, we use the example of the second and fifth variables, which were significantly correlated with each other, and compare how the HNB and TLNPN models modeled their joint distribution as compared with the test data. We see in **Figure 17** that there is a dependence between the variables in the test data, which the TLNPN data captures, but the HNB data does not, leading to a higher Wasserstein distance for the HNB simulated data.

Single Cell RNA Sequencing Data

As another example of real data analysis, we use single-cell RNA sequencing data from the lymphoblastoid cell line; the original data can be found on the 10x Genomics Datasets website (*https://www.10xgenomics.com*). This dataset measures $p = 329$ genes from $n = 265$ cells. The first, second, third, and fourth quartiles of the zero-proportion of the variables in the RNA dataset are 41.1%, 65.7%, 81.1%, and 89.8% respectively. In this case, we did not have to transform the data as all data points were already well below 2^{31} and 95% are below 10. However, we found in **Figure 18** that the TLNPN model and HNB model performed similarly over the 50 replications for the p-dimensional Wasserstein distance. Furthermore, we found that the HNB and TLNPN models performed similarly modeling the marginal distributions of the p variables.

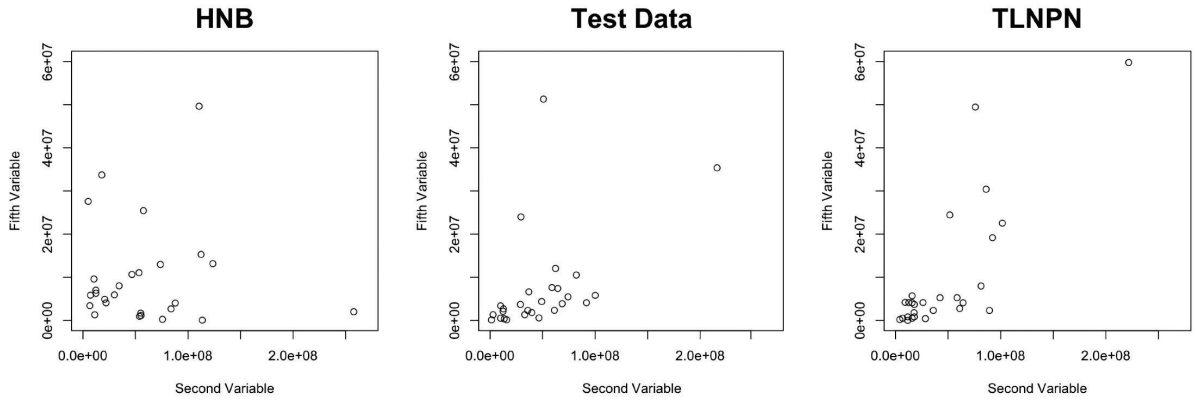

Figure 17. Here, we compare the joint distributions of second and fifth variables of the HNB, TLNPN, and QMP test data to show how the TLNPN model is able to model the dependence of the test data that the HNB model cannot.

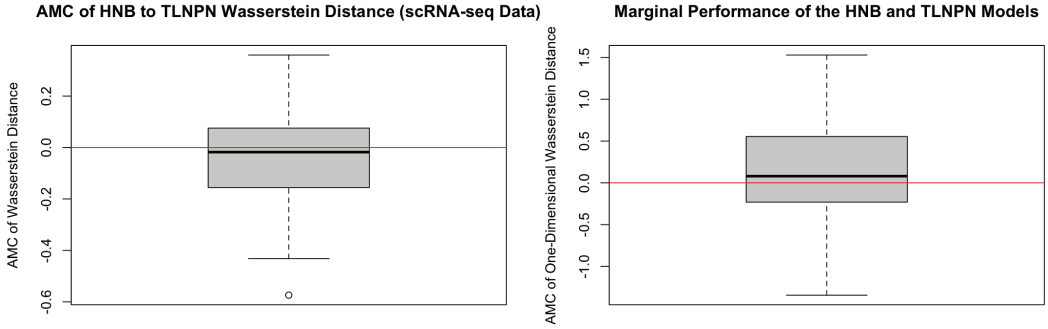

Figure 18. The first box plot compares the overall performance of the TLNPN model to the HNB model under the RNA sequencing data. The y-axis displays the AMC of the Wasserstein distance generated by the HNB model to that of the TLNPN model, defined in **Equation 3**. The results are based on 50 random splits of the gene sequencing data. The red line at zero marks the reference: points above indicate better HNB fit, and points below indicate better TLNPN fit. The second box plot shows how the models compared when modeling marginal distributions. Overall, the HNB and TLNPN models performed similarly both for the joint distribution and the marginal distributions.

In this real-data analysis, we compared the TLNPN model to the HNB model using single cell RNA sequencing data; the results of which are displayed in **Figure 18**. We see that the models perform similarly on a multivariate and marginal level. Based on these results, we conjecture that the characteristics of this dataset, the small scale and a lack of highly correlated variables, resulted in the similar performance of the HNB and TLNPN models. The TLNPN model is more robust to extreme values than the HNB model fitted without covariates, but this dataset had very little skewness and variance in comparison to the QMP dataset, which contributed to the TLNPN model performing similarly to the HNB model.

CONCLUSION

In this work, we sought to compare models for zero-inflated data through both simulation and real data studies that mimicked and used modern biomedical data. Zero-inflated and hurdles models have been popularly used in this field and we sought to compare them with the newly introduced truncated latent Gaussian copula model. The recent emergence of the TLNPN model created a gap in the literature comparing the TLNPN model to the established zero-inflated and hurdle models, and this paper sought to compare these models under different circumstances. Furthermore, in this work, we sought to find the weaknesses of the TLNPN model such as its underestimation of correlation among zero-inflated variables and its struggles in modeling marginal distributions. We found in the simulation studies and real data analyses that the main considerations for deciding to fit either the TLNPN or the HNB models were access to covariates, variance of the data, and dependence among the variables.

The obvious advantage of using the TLNPN model is that it can account for dependence among variables without having access to

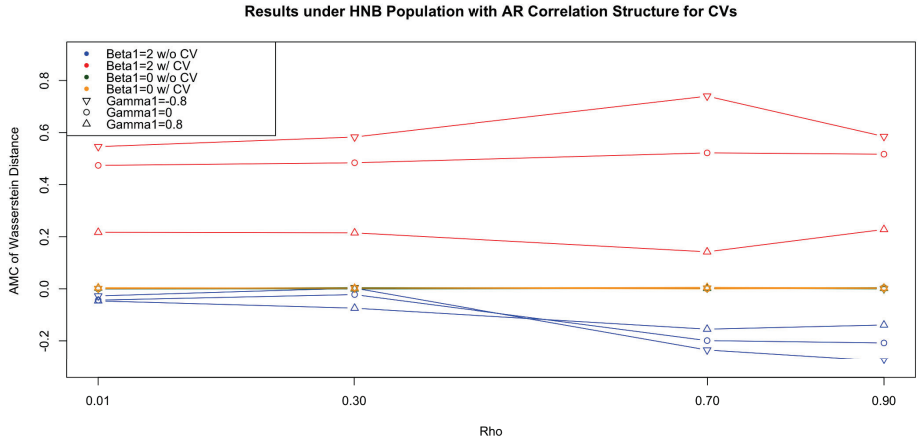

Figure 19. We summarize the results of Simulation Setting Two under HNB population data where the covariates follow an AR correlation structure. We see that when $\beta_1 = 2$, the HNB model fitted with covariates outperforms the TLNPN model; as γ_1 decreases, the AMC further increases. When the HNB model is fitted without covariates, the opposite trend emerges where the TLNPN model outperforms the HNB model, and the AMC decreases as γ_1 decreases (when $\rho = 0.7$ or $\rho = 0.9$). We also see that when $\beta_1 = 0$, the two models perform nearly identically regardless of ρ, γ_1, or whether covariates are fitted in the HNB model.

the covariates in contrast to the HNB model. However, the TLNPN requires a large amount of training data to accurately estimate the correlation of the Gaussian-level variables, and furthermore, in cases of strong dependence among the variables, the TLNPN model tends to underestimate the correlation between the zero-inflated variables when the true model is a HNB model. Furthermore, when the HNB model has access to the covariates, it tends to model the dependence structure between the variables much more accurately. Nevertheless, when no covariate is available, the TLNPN model typically outperforms the HNB model in fitting multivariate distributions of highly dependent zero-inflated variables. We see in **Figure 19** where the HNB model fitted with covariates outperforms the TLNPN model when $\beta_1 = 2$ on account of its ability to better model dependence among the zero-inflated variables and the marginal distributions. However, when $\beta_1 = 2$ and the HNB model is fitted without covariates, the TLNPN model outperforms the HNB model.

Another drawback of the TLNPN model is that when the true population is HNB with a high β_1 parameter, the TLNPN model struggles to model large, outlier values compared to the HNB model fitted with covariates. The TLNPN model will never predict a data point outside of its original training data because of its use of the empirical CDF, so the training data must be similar to the testing data for it to perform well. However, on a marginal level, the HNB model itself can be vulnerable to overdispersed and highly skewed data, which the TLNPN model is better at fitting marginally. Furthermore, a computational limitation of the HNB model is that the main function used to fit the data to a HNB model can only handle integer values below 2^{31}, so datasets with large values may need to be rescaled, as modern biomedical datasets often contain extremely large measurements. This can pose challenges for statistical analysis, as results may vary depending on the rescaling method used.

Future research could investigate how the TLNPN model performs against other models, particularly zero-inflated Poisson and hurdle Poisson models with an overdispersion parameter. Furthermore, investigation of incorporating covariates into the TLNPN model will be an interesting research direction to pursue.

ACKNOWLEDGEMENTS

This research was supported by the 2024 Mathematics Research Experiences for Undergraduates (REU) program at the University of North Carolina at Charlotte under NSF-REU grant DMS-2150179.

REFERENCES

1. Feng, C. X. (2021) A comparison of zero-inflated and hurdle models for modeling zero-inflated count data, *J Stat Distrib App* 8, 8. *https://doi.org/10.1186/s40488-021-00121-4*
2. Vandeputte, D., Kathagen, G. and D'hoe, K., and Vieira-Silva, S., and Valles-Colomer, M., and Sabino, J., and Wang, J., and Tito, R. Y., and De Commer, L., and Darzi, Y., Vermeire, S., Falony, G., and Raes, J. (2017) Quantitative microbiome profiling links gut community variation to microbial load, *Nature* 551, 507–511. *https://doi.org/10.1038/nature24460*
3. Li, H. (2015) Microbiome, metagenomics, and high-dimensional compositional data analysis, *Annu. Rev. Stat. Appl.* 2, 73–94. *https://doi.org/10.1146/annurev-statistics-010814-020351*
4. Peruman-Chaney, S., Morgan, C., McDowall, D., and Aban, I., (2013) Zero-inflated and overdispersed: what's one to do?, *JSCS* 2, 59–67. *https://doi.org/10.1080/00949655.2012.668550*
5. Freckleton, R. P., (2009) The seven deadly sins of comparative analysis, *J. Evol. Biol.* 22, 1367–1375. *https://doi.org/10.1111/j.1420-9101.2009.01757.x*
6. Campbell, H. (2021) The consequences of checking for zero-inflation and overdispersion in the analysis of count data, *MEE* 12, 665–680. *https://doi.org/10.1111/2041-210X.13559*
7. Hua, H., Wan, T., and Crits-Christoph, P. (2014) Structural zeroes and zero-inflated models, *Shanghai Arch. Psychiatry* 26, 236–242. *10.3969/j.issn.1002-0829.2014.04.008*
8. Rose, C. E., Martin, S. W., Wannemuehler, K. A., and Plikaytis., B. D. (2006) On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data, *J. Biopharm. Stat.* 16, 463–481. *https://doi.org/10.1080/10543400600719384*
9. Dong, C., Clarke, D., Yan, X., Khattak, A., and Huang, B. (2014) Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections, *Accident Analysis and Prevention* 70, 320–329. *https://doi.org/10.1016/j.aap.2014.04.018*
10. Yoon, G., Carroll, R. J., and Gaynanova, I. (2020) Sparse semiparametric canonical correlation analysis for data of mixed types, *Biometrika* 107, 609–625. *https://doi.org/10.1093/biomet/asaa007*
11. Fan, J., Liu, H., Ning, Y., and Zou, Hui (2017) High dimensional semiparametric latent graphical model for mixed data, *J. R. Stat. Soc. Ser. B Methodol.* 79, 405–421. *https://doi.org/10.1111/rssb.12168*
12. Saffari, S. E., Adnan, R., and Greene, W. (2012) Hurdle negative binomial regression model with right censored count data, *SORT* 36, 181–194.
13. Liu H., Lafferty J., and Wasserman L. (2009) The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs, *Journal of Machine Learning Research* 10, 2295–2328. *https://www.jmlr.org/papers/volume10/liu09a/liu09a.pdf*
14. Chung H. C., Ni, Y., and Gaynanova, I. (2022) Sparse semiparametric discriminant analysis for high-dimensional zero-inflated data, *arXiv*. *https://doi.org/10.48550/arXiv.2208.03734*
15. Panaretos, V. and Zemel, Y. (2018) Statistical Aspects of Wasserstein Distances, *Annu. Rev. Stat. Appl.* 6, 401–431. *https://doi.org/10.1146/annurev-statistics-030718-104938*
16. Schuhmacher D, Bähre B, Bonneel N, Gottschlich C, Hartmann V, Heinemann F, Schmitzer B, Schrieber J (2024). transport: Computation of Optimal Transport Plans and Wasserstein Distances. R package version 0.15-4, https://cran.r-project.org/package=transport.
17. Chikuse, Y. (2012). Statistics on special manifolds (Vol. 174). Springer Science & Business Media.
18. Zeileis, A., Kleiber, C., and Jackman, S. (2008) Regression Models for Count Data in R. *J. Stat. Softw.*, 27, 1–25. *https://doi.org/10.18637/jss.v027.i08*

ABOUT THE STUDENT AUTHORS

Max Beveridge will graduate from the George Washington University in May of 2025 with a Bachelor of Science in Statistics and International Affairs. Zachary Goldstein will graduate from Yeshiva University with a Bachelor of Arts in Pure/Applied Mathematics and Mathematical Economics in May of 2026.

PRESS SUMMARY

Many modern biomedical datasets have variables that are zero-inflated, and modeling these zeros correctly is critical for accurate statistical analysis. We evaluate three models (zero-inflated negative binomial, hurdle negative binomial, and the truncated latent Gaussian copula models) to see which performs the best under varying conditions. Specifically, we seek to evaluate whether the level of dependence among the variables impacts which model performs the best.