A Flexible Quasi-Copula Distribution for Statistical Modeling

Sarah S. Ji\textsuperscript{1}\textsuperscript{†}, Benjamin B. Chu\textsuperscript{2}, Janet S. Sinsheimer\textsuperscript{1,3,4}, Hua Zhou\textsuperscript{1,4}, Kenneth Lange\textsuperscript{3,4,5}

\textsuperscript{1}Department of Biostatistics, University of California, Los Angeles, CA
\textsuperscript{2}Department of Biomedical Data Science, Stanford University, Stanford, CA
\textsuperscript{3}Department of Human Genetics, University of California, Los Angeles, CA
\textsuperscript{4}Department of Computational Medicine, University of California, Los Angeles, CA
\textsuperscript{5}Department of Statistics, University of California, Los Angeles, CA

May 10, 2022

Abstract

Copulas, generalized estimating equations, and generalized linear mixed models promote the analysis of grouped data where non-normal responses are correlated. Unfortunately, parameter estimation remains challenging in these three frameworks. Based on prior work of Tonda, we derive a new class of probability density functions that allow explicit calculation of moments, marginal and conditional distributions, and the score and observed information needed in maximum likelihood estimation. Unlike true copulas, our quasi-copula model only approximately preserves marginal distributions. Simulation studies with Poisson, negative binomial, Bernoulli, and Gaussian bases demonstrate the computational and statistical virtues of the quasi-copula model and its limitations.

\textit{Keywords:} Linear mixed models, MM principle, maximum likelihood

\textsuperscript{†}The authors gratefully acknowledge USPHS, Grant Number: GM141798, AI153044, HG002536 and HG006139. NSF, Grant Number: DMS-2054253
# Contents

1 Introduction 4

2 Definitions 5

3 Moments 6

4 Marginal and Conditional Distributions 8

5 Generation of Random Deviates 9

6 Parameter Estimation 10
   6.1 Mean Components . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
   6.2 Structured Covariance . . . . . . . . . . . . . . . . . . . . . . . . . 11
   6.3 MM Algorithm for the VC Model Parameters . . . . . . . . . . . . . 12
   6.4 Initialization . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

7 Statistical Properties 13
   7.1 Compound Symmetric Covariance . . . . . . . . . . . . . . . . . . . . 14

8 Results 14
   8.1 Simulation Studies . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
      8.1.1 Run Times . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
   8.2 Bivariate Mixed Outcome Model . . . . . . . . . . . . . . . . . . . . 24
   8.3 NHANES Data Example . . . . . . . . . . . . . . . . . . . . . . . . . 24

9 Discussion 26

10 Web Resources 27

11 Supplemental Materials 30
   11.1 Tonda’s Approximation Details . . . . . . . . . . . . . . . . . . . . 30
   11.2 Generate Random Deviates . . . . . . . . . . . . . . . . . . . . . . . 30
      11.2.1 Marginal Distribution . . . . . . . . . . . . . . . . . . . . . . 31
      11.2.2 Conditional Distribution . . . . . . . . . . . . . . . . . . . . . 33
      11.2.3 Continuous Outcomes . . . . . . . . . . . . . . . . . . . . . . 34
      11.2.4 Discrete Outcomes . . . . . . . . . . . . . . . . . . . . . . . . 39
   11.3 Parameter Estimation: . . . . . . . . . . . . . . . . . . . . . . . . . . 43
      11.3.1 Fisher Scoring to Estimate Beta . . . . . . . . . . . . . . . . . 43
11.3.2 MM-Algorithm to Update Variance Components . . . . . . . . 46
11.3.3 Quasi-Newton Algorithm . . . . . . . . . . . . . . . . . . . . 47
11.3.4 Negative Binomial . . . . . . . . . . . . . . . . . . . . . . . . 48
11.3.5 Compound Symmetric Covariance: . . . . . . . . . . . . . . . 51
11.4 Gaussian Base . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
11.4.1 Moments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
11.4.2 Parameter Estimation . . . . . . . . . . . . . . . . . . . . . . 54
11.5 Additional Simulation Study Results . . . . . . . . . . . . . . . . 58
11.5.1 AR(1) Covariance . . . . . . . . . . . . . . . . . . . . . . . . . 60
11.5.2 CS Covariance . . . . . . . . . . . . . . . . . . . . . . . . . . 64
11.5.3 VC Covariance . . . . . . . . . . . . . . . . . . . . . . . . . . 68
11.5.4 Run Times . . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
1 Introduction

The analysis of correlated data is stymied by the lack of flexible multivariate distributions with fixed margins. Once one ventures beyond the confines of multivariate Gaussian distributions, analysis choices are limited. Liang and Zeger (1986) launched the highly influential method of generalized estimating equations (GEEs). This advance allows generalized linear models (GLMs) to accommodate the correlated traits encountered in panel and longitudinal data and effectively broke the stranglehold of Gaussian distributions in analysis. The competing method of statistical copulas introduced earlier by Sklar is motivated by the same consideration (Sklar, 1959). Finally, generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993; Zeger and Karim, 1991) attacked the same problem. GLMMs are effective tools for modeling overdispersion and capturing the correlations of multivariate discrete data.

However, none of these three modeling approaches is a panacea. GEEs lack a well-defined likelihood, and estimation searches can fail to converge. For copula models, likelihoods exist, but are unwieldy, particularly for discrete outcomes. Copula calculations scale extremely poorly in high dimensions. Computing with GLMMs is problematic since their densities have no closed form and require evaluation of multidimensional integrals. Gaussian quadratures scale exponentially in the dimension of the parameter space. Markov Chain Monte Carlo (MCMC) can be harnessed in Bayesian versions of GLMMs, but even MCMC can be costly. For these reasons alone, it is worth pursuing alternative modeling approaches.

This brings us to an obscure paper by the Japanese mathematical statistician Tonda. Working within the framework of Gaussian copulas (Song et al., 2009) and generalized linear models, Tonda introduces a device for relaxing independence assumptions while preserving computable likelihoods (Tonda, 2005). He succeeds brilliantly except for the presence of an annoying constraint on the parameter space of the new distribution class. The fact that his construction perturbs marginal distributions is forgivable. The current paper has several purposes. First, by adopting a slightly different working definition, we show how to extend his construction to lift the awkward parameter constraint. Our new definition allows explicit calculation of (a) moments, (b) marginal and conditional distributions, and (c) the score and observed information of the loglikelihood and allows (d) generation of random deviates. Tonda tackles item (a), omits items (b) and (c), and mentions item (d) only in passing. For maximum likelihood estimation (MLE), he relies on a non-standard derivative-free algorithm (Ohtaki, 1999) that scales poorly in high dimensions. We present two gradient-based algorithms designed for high-dimensional MLEs. The first is a block ascent algorithm that updates fixed effects by Newton’s method and
updates variance components by a minorization-maximization (MM) algorithm. The second is a standard quasi-Newton algorithm that updates fixed effects and variance components jointly.

In contrast to other multivariate outcome models, our loglikelihoods contain no determinants or matrix inverses. These features resolve computational bottlenecks in parameter estimation. We advocate gradient based estimation methods that avoid computationally intensive second derivatives. Approximate Hessians can be computed after estimation to provide asymptotic standard errors and confidence intervals. The range of potential applications of our quasi-copula model is enormous. Panel, longitudinal, time series, and all of GLM modeling stand to benefit. In addition to relaxing independence assumptions, our models offer a simple way to capture over-dispersion. Our simulation studies and real data examples highlight not only the virtues of the quasi-copula model but also its limitations. For reasons to be explained, we find that the model reflects reality best when the size of the independent sampling units is low or the correlations between responses within a unit are small.

2 Definitions

Consider \(d\) independent random variables \(X_1, \ldots, X_d\) with densities \(f_i(x_i)\) relative to measures \(\alpha_i\), with means \(\mu_i\), variances \(\sigma_i^2\), third central moments \(c_i^3\), and fourth central moments \(c_i^4\). Let \(\Gamma = (\gamma_{ij})\) be an \(d \times d\) positive semidefinite matrix, and \(\alpha\) be the product measure \(\alpha_1 \times \cdots \times \alpha_d\). Inspired by Tonda (2005), we let \(D\) be the diagonal matrix with \(i\)th diagonal entry \(\sigma_i\) and consider the nonnegative function

\[
1 + \frac{1}{2} (x - \mu)^t D^{-1} \Gamma D^{-1} (x - \mu).
\]

Its average value is

\[
\int \prod_{i=1}^{d} f_i(x_i) \left[ 1 + \frac{1}{2} (x - \mu)^t D^{-1} \Gamma D^{-1} (x - \mu) \right] d\alpha(x) = 1 + \frac{1}{2} \sum_i \sum_j E \left[ \frac{(x_i - \mu_i)(x_j - \mu_j)}{\sigma_i \sigma_j} \right] \gamma_{ij}
\]

\[
= 1 + \frac{1}{2} \sum_i \gamma_{ii}.
\]

It follows that the function

\[
g(x) = \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \prod_{i=1}^{d} f_i(x_i) \left[ 1 + \frac{1}{2} (x - \mu)^t D^{-1} \Gamma D^{-1} (x - \mu) \right]
\]

(1)
is a probability density with respect to the measure \( \alpha \). Detailed derivations of Tonda’s Approximation are found in the supplemental material. The virtue of the density is that it overcomes the independence restriction and steers the sample matrix of the residuals toward the target covariance matrix \( \Gamma \). Note that \( g(x) \) is technically not a copula since it fails to preserve the marginal distributions \( f_i(x_i) \). For instance, we will see later that \( g(x) \) tends to inflate marginal variances.

3 Moments

Let \( Y = (Y_1, \ldots, Y_d)^t \) be a random vector distributed as \( g(x) \). To calculate the mean of \( Y_k \), note that our independence assumption implies

\[
\int (x_k - \mu_k) g(x) \alpha(x) = \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \frac{1}{2} \sum_i \sum_j \mathbb{E} \left[ (x_k - \mu_k) (x_i - \mu_i) (x_j - \mu_j) \right] \gamma_{ij} \\
= \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \frac{\kappa_{k3}}{2\sigma_k^2}.
\]

Hence, if \( \kappa_{k3} \) is the skewness of \( X_k \), then

\[
\mathbb{E}(Y_k) = \mu_k + \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \frac{\kappa_{k3} \gamma_{kk}}{2\sigma_k^2} \\
= \mu_k + \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \frac{\sigma_k K_{k3} \gamma_{kk}}{2} \\
= \mu_k + \frac{\sigma_k K_{k3} \gamma_{kk}}{2} + O(\|\Gamma\|^2)
\]

for any matrix norm \( \|\Gamma\| \). The mean \( \mathbb{E}(Y_k) \) is close to \( \mu_k \) when the diagonal entries of \( \Gamma \) and, hence \( \|\Gamma\| \) itself, are small.

To calculate the covariance matrix of \( Y \), note that

\[
\int (x_k - \mu_k) (x_l - \mu_l) g(x) d\alpha(x) = \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} 1_{(k=l)} \sigma_k^2 + \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \times \frac{1}{2} \sum_i \sum_j \mathbb{E} \left[ (x_k - \mu_k) (x_i - \mu_i) (x_j - \mu_j) \right] \gamma_{ij}.
\]
The indicated expectations relative to $\prod_{i=1}^{d} f_i(x_i)$ reduce to

$$E[(x_k - \mu_k)(x_l - \mu_l)(x_i - \mu_i)(x_j - \mu_j)]$$

$$= \begin{cases} c_{k4} & k = l = i = j \\ \sigma_k^2 \sigma_l^2 & k = l \neq i = j \\ \sigma_k^2 \sigma_i^2 & k = i \neq l = j \\ \sigma_k^2 \sigma_j^2 & k = j \neq l = i \\ 0 & \text{otherwise} \end{cases}$$

When $k = l$ and $\kappa_{k4}$ is the kurtosis of $X_k$,

$$\int (x_k - \mu_k)^2 g(x) d\alpha(x) = \left[1 + \frac{1}{2} \text{tr}(\Gamma)\right]^{-1} \left[\sigma_k^2 + \frac{1}{2} c_{k4} \gamma_{kk} + \frac{1}{2} \sigma_k^2 \sum_{i \neq k} \gamma_{ii}\right]$$

$$= \left[1 + \frac{1}{2} \text{tr}(\Gamma)\right]^{-1} \sigma_k^2 \left[1 + \frac{\kappa_{k4} \gamma_{kk}}{2} + \frac{1}{2} \sum_{i \neq k} \gamma_{ii}\right]$$

$$= \sigma_k^2 + \frac{\sigma_k^2 \kappa_{k4} \gamma_{kk}}{2} + \frac{\sigma_k^2}{2} \sum_{i \neq k} \gamma_{ii} + O(\|\Gamma\|^2)$$

$$= \sigma_k^2 \left[1 + \frac{(\kappa_{k4} - 1) \gamma_{kk}}{2} + \frac{1}{2} \sum_i \gamma_{ii}\right] + O(\|\Gamma\|^2)$$

Because $[E(Y_k - \mu_k)]^2 = O(\|\Gamma\|^2)$, we find that

$$\text{Var}(Y_k) = E[(Y_k - \mu_k)^2] = [E(Y_k - \mu_k)]^2$$

$$= \sigma_k^2 \left[1 + \frac{(\kappa_{k4} - 1) \gamma_{kk}}{2} + O(\|\Gamma\|^2)\right]$$

Because the kurtosis $\kappa_{k4} \geq 1$, the multiplier $\kappa_{k4} - 1$ of $\gamma_{kk}$ is nonnegative, and the variance is inflated for $\|\Gamma\|$ small.

When $k \neq l$,

$$\int (x_k - \mu_k)(x_l - \mu_l) g(x) d\alpha(x) = \left[1 + \frac{1}{2} \text{tr}(\Gamma)\right]^{-1} \frac{1}{2} \sigma_k \sigma_l \gamma_{kl}$$

$$= \sigma_k \sigma_l \gamma_{kl} + O(\|\Gamma\|^2).$$
Hence, the covariance and correlation satisfy
\[
\begin{align*}
\text{Cov}(Y_k, Y_l) &= \text{Cov}(Y_k - \mu_k, Y_l - \mu_l) \\
&= E[(Y_k - \mu_k)(Y_l - \mu_l)] - E(Y_k - \mu_k)E(Y_l - \mu_l) \\
&= \sigma_k \sigma_l \gamma_{kl} + O(\|\Gamma\|^2)
\end{align*}
\]
\[
\text{Corr}(Y_k, Y_l) = \frac{\gamma_{kl}}{\sqrt{1 + \frac{(\kappa_k - 1)\gamma_{kk}}{2} + O(\|\Gamma\|^2)}}
\]

As a check, the quantities \(E(Y_k), \text{Var}(Y_k), \text{Cov}(Y_k, Y_l)\) reduce to the correct values \(\mu_k, \sigma_k^2, \text{and } 0\), respectively, when \(\Gamma = 0\).

4 Marginal and Conditional Distributions

Let \(S\) be a subset of \(\{1, \ldots, d\}\) with complement \(T\). To simplify notation, suppose \(S = \{1, 2, \ldots, s\}\). Now write
\[
Y = \begin{pmatrix} Y_S \\ Y_T \end{pmatrix}, \quad r = \begin{pmatrix} r_S \\ r_T \end{pmatrix}, \quad \Gamma = \begin{pmatrix} \Gamma_S & \Gamma_{ST} \\ \Gamma_{ST}^T & \Gamma_T \end{pmatrix}, \quad \alpha = \alpha_S \times \alpha_T,
\]
where \(r\) is the vector \(D^{-1}(Y - \mu)\) of standardized residuals. The marginal density of \(Y_S\) is
\[
\begin{align*}
&\left[1 + \frac{1}{2} \text{tr}(\Gamma)\right]^{-1} \prod_{i \in S} f_i(y_i) \int \prod_{i \in T} f_i(y_i) \left[1 + \frac{1}{2} r^T \Gamma r\right] d\alpha_T(y_T) \\
&= \left[1 + \frac{1}{2} \text{tr}(\Gamma)\right]^{-1} \prod_{i \in S} f_i(y_i) \left[1 + \frac{1}{2} r_s^T \Gamma_S r_s + \frac{1}{2} \text{tr}(\Gamma_T)\right].
\end{align*}
\]

To derive the conditional density of \(Y_S\) given by \(Y_T\), we divide the joint density by the marginal density of \(Y_T\). This action produces the conditional density
\[
d_S \prod_{i \in S} f_i(y_i) \left[1 + \frac{1}{2} r^T \Gamma r\right]
\]
with normalizing constant \(d_S = \left[1 + \frac{1}{2} r_s^T \Gamma_T r_T + \frac{1}{2} \text{tr}(\Gamma_S)\right]^{-1}\). From this density, our well-rehearsed arguments lead to the conditional mean
\[
E(Y_k | Y_T) = \mu_k + d_S \left[\frac{ck_3 \gamma_{kk}}{2\sigma_k^2} + \frac{1}{\sigma_k} \sum_{j \in T} r_j \gamma_{jk}\right] = \mu_k + \frac{ck_3 \gamma_{kk}}{2\sigma_k^2} + O(\|\Gamma\|^2)
\]
for \( k \in S \). The corresponding conditional variance is

\[
\text{Var}(Y_k \mid Y_T) = \sigma_k^2 + \frac{1}{2} \left( \frac{c_{k1}}{\sigma_k^2} - \sigma_k^2 \right) \gamma_{kk} + \sum_{j \in T} \frac{c_{k3j} \gamma_{kj}}{\sigma_k} + O(\|\Gamma\|^2).
\]

and the corresponding conditional covariances are

\[
\text{Cov}(Y_k, Y_l \mid Y_T) = \sigma_k \sigma_l \gamma_{kl} + O(\|\Gamma\|^2)
\]

for \( k \in S, l \in S, \) and \( k \neq l \). It is noteworthy that to order \( O(\|\Gamma\|^2) \), the conditional and marginal means agree, and the conditional and marginal covariances agree.

## 5 Generation of Random Deviates

To generate a random vector from the density (1), we first sample \( Y_1 \) from its marginal density

\[
\left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} f_1(y_1) \left( 1 + \frac{\gamma_{11}}{2} r_1^2 + \frac{1}{2} \sum_{i=2}^{d} \gamma_{ii} \right),
\]

and then sample the subsequent components \( Y_i \) from their conditional distributions \( Y_i \mid Y_1, \ldots, Y_{i-1}, \forall i \in [1, d] \). If we denote the set \( \{1, \ldots, i-1\} \) by \( [i-1] \), then the conditional density of \( Y_i \) given the previous components is

\[
d_{[i-1]}^{-1} f_i(y_i) \left[ d_{[i-1]} + r_i \sum_{j=1}^{i-1} r_j \gamma_{ij} + \frac{\gamma_{ii}}{2} (r_i^2 - 1) \right],
\]

where \( d_{[i-1]} = 1 + \frac{1}{2} r_{[i-1]} \Gamma_{[i-1]} r_{[i-1]} + \frac{1}{2} \sum_{j=1}^{d} \gamma_{jj} \).

When the densities \( f_i(y_i) \) are discrete, each stage of sampling is straightforward. Consider any random variable \( Z \) with nonnegative integer values, discrete density \( p_i = \Pr(Z = i) \), and mean \( \nu \). The inverse method of random sampling reduces to a sequence of comparisons. We partition the interval \([0, 1]\) into subintervals with the \( i \)th subinterval of length \( p_i \). To sample \( Z \), we draw a uniform random deviate \( U \) from \([0, 1]\) and return the deviate \( j \) determined by the conditions \( \sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^{j} p_i \). The process is most efficient when the largest \( p_i \) occur first. This suggests that we let \( k \) denote the least integer \( \lceil \nu \rceil \) and rearrange the probabilities in the order \( p_k, p_{k+1}, p_{k-1}, p_{k+2}, p_{k-2}, \ldots \). This tactic is apt put most of the probability mass first and render sampling efficient.
When the densities \( f_i(y_i) \) are continuous, each stage of sampling is probably best performed by inverse transform sampling. This requires calculating distribution functions and forming their inverses, either analytically or by Newton’s method. The required distribution functions assume the form

\[
\int_{-\infty}^{x} f(y)[a_0 + a_1(y - \mu) + a_2(y - \mu)^2] dy = \int_{-\infty}^{x} f(y)[b_0 + b_1y + b_2y^2] dy.
\]

The integrals \( \int_{-\infty}^{x} f(y)y^j dy \) are available as special functions for Gaussian, beta, and gamma densities \( f(y) \). For instance, if \( \phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \) is the standard normal density and \( \Phi(x) \) is the standard normal distribution, then

\[
\int_{-\infty}^{x} y\phi(y) dy = -\phi(x) \quad \text{and} \quad \int_{-\infty}^{x} y^2\phi(y) dy = \Phi(x) - x\phi(x).
\]

To avoid overburdening the text with classical mathematics, we omit further details. Additional derivations can be found in the supplemental material.

6 Parameter Estimation

6.1 Mean Components

Consider \( n \) independent realizations \( y_i \) from the quasi-copula density (1). Each of these may be of a different dimension \( d_i \) and possess a different mean vector \( \mu_i(\beta) \), covariance matrix \( \Gamma_i(\theta) = \gamma_{ijk}(\theta) \), and component densities \( f_{ij}(y_{ij} | \beta) \). If \( r_i(\beta) \) denotes the vector \( D_i^{-1}(y_i - \mu_i) \) of standardized residuals for sampling unit \( i \), then the loglikelihood of the sample is

\[
\mathcal{L}(\beta, \theta) = -\sum_{i=1}^{n} \ln \left[ 1 + \frac{1}{2} \text{tr}(\Gamma_i(\theta)) \right] + \sum_{i=1}^{n} \sum_{j=1}^{d_i} \ln f_{ij}(y_{ij} | \beta) + \sum_{i=1}^{n} \ln \left\{ 1 + \frac{1}{2} r_i(\beta)^T \Gamma_i(\theta) r_i(\beta) \right\}.
\]

The score (gradient of the loglikelihood) with respect to \( \beta \) is clearly

\[
\nabla_{\beta} \mathcal{L}(\beta, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{d_i} \nabla \ln f_{ij}(y_{ij} | \beta) + \sum_{i=1}^{n} \frac{\nabla r_i(\beta)^T \Gamma_i(\theta) r_i(\beta)}{1 + \frac{1}{2} r_i(\beta)^T \Gamma_i(\theta) r_i(\beta)}.
\]
where $\nabla r_i(\beta)^t = dr_i(\beta)$ is the differential (Jacobi matrix) of the vector $r_i(\beta)$. An easy calculation shows that $\nabla r_i(\beta)$ has entries

$$\nabla_{r_{ij}}(\beta) = -\frac{1}{\sigma_{ij}(\beta)} \nabla \mu_{ij}(\beta) - \frac{1}{2} \frac{\nu_{ij}(\beta)}{\sigma_{ij}^3(\beta)} \nabla \sigma_{ij}^2(\beta).$$

In searching the likelihood surface, it is best to approximate the observed information by a positive definite matrix. This suggests replacing $-d^2 \ln f_{ij}(y_{ij} | \beta)$ by the expected information matrix $J_{ij}(\beta)$ under the independence model and dropping indefinite matrices in the exact Hessian. These steps give the approximate Hessian

$$d^2 \beta L \approx -\sum_{i=1}^n \sum_{j=1}^{d_i} J_{ij}(\beta) - \sum_{i=1}^n \left[ \nabla r_i(\beta) \Gamma_i(\theta) r_i(\beta) \right]^t \left[ 1 + \frac{1}{2} r_i(\beta)^t \Gamma_i(\theta) r_i(\beta) \right]^2,$$

which is clearly negative semidefinite. As partial justification for this approximation, we expect residuals to be small on average. The score and approximate Hessian provide the ingredients for an approximate Newton’s method for improving $\beta$.

### 6.2 Structured Covariance

Maximization of the loglikelihood also involves finding optimal values for the covariance parameters $\theta$ determining the structured covariance matrices $\Gamma_i$. Assuming there are no shared mean and covariance parameters, the relevant part of the loglikelihood is

$$-\sum_{i=1}^n \ln \left[ 1 + \frac{1}{2} \text{tr}(\Gamma_i(\theta)) \right] + \sum_{i=1}^n \ln \left[ 1 + \frac{1}{2} r_i(\beta)^t \Gamma_i(\theta) r_i(\beta) \right].$$

To simplify estimation of $\Gamma_i$, we investigate just three covariance scenarios, namely, an autoregressive AR(1) model, a compound symmetric (CS) model, and a variance components (VC) model. Under the AR(1) and CS models, $\Gamma_i(\theta)$ is parameterized by $\theta = (\sigma^2, \rho)$, a total variance $\sigma^2$ and a correlation $\rho$. For the AR(1) model this leads to the representation

$$\Gamma_i(\theta) = \sigma^2 \times \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \ldots & \rho^{d_i-1} \\ \rho & 1 & \rho & \rho^2 & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ \rho^{d_i-1} & \rho^{d_i-2} & \ldots & \rho^2 & \rho & 1 \end{bmatrix}$$

$$= \sigma^2 \times V_i(\rho).$$
For the CS model this leads to the representation
\[ \Gamma_i = \sigma^2 \times \left( \rho 1_{d_i} 1_{d_i}^t + (1 - \rho) I_{d_i} \right) \]
\[ = \sigma^2 \times V_i(\rho) \]

The relevant part of the loglikelihood can be rewritten as
\[ f(\sigma^2, \rho) = -\sum_{i=1}^{n} \ln \left[ 1 + \frac{d_i \sigma^2}{2} \right] + \sum_{i=1}^{n} \ln \left[ 1 + \frac{\sigma^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta) \right] \]

A typical variance components problem depends on the decomposition \( \Gamma_i(\theta) = \sum_{k=1}^{m} \theta_k \Omega_{ik} \) of \( \Gamma_i \) into a linear combination of known covariance matrices \( \Omega_{ik} = (\omega_{ikjl}) \) against unknown nonnegative variance components \( \theta_k \) arranged in a vector \( \theta = (\theta_k) \). Now the relevant part of the loglikelihood amounts to
\[ f(\theta) = \sum_{i=1}^{n} \ln(1 + \theta^t b_i) - \sum_{i=1}^{n} \ln(1 + \theta^t c_i), \]
where the vectors \( b_i \) and \( c_i \) have the nonnegative components
\[ b_{ik} = \frac{1}{2} r_i(\beta)^t \Omega_{ik} r_i(\beta) \text{ and } c_{ik} = \frac{1}{2} \text{tr}(\Omega_{ik}). \]

Derivation of the scores \( \nabla_{\theta} \mathcal{L} \) and approximate observed information matrices \( -d_{\theta}^2 \mathcal{L} \) for the AR(1), CS and VC models is relegated to the Supplemental Material. The gradients alone provide the raw material for a quasi-Newton search of parameter space.

### 6.3 MM Algorithm for the VC Model Parameters

One can construct an iterative MM algorithm for updating \( \theta \) holding \( \beta \) fixed. There exists a substantial literature on the MM principle for optimization (Lange et al., 2000; Lange, 2016; Zhou et al., 2019). The idea in maximization is to concoct a surrogate function \( g(\theta \mid \theta_r) \) that is easy to maximize and hugs the objective \( f(\theta) \) tightly. Here \( \theta_r \) is the current value of \( \theta \). Construction of the surrogate is guided by two minorization requirements:
\[ f(\theta) \geq g(\theta \mid \theta_r) \quad \forall \theta \text{ (dominance condition)} \]
\[ f(\theta_r) = g(\theta_r \mid \theta_r) \text{ (tangent condition)}. \]
The next iterate is determined by \( \theta_{r+1} = \text{argmax} \ g(\theta \mid \theta_r) \). The MM principle guarantees that \( f(\theta_{r+1}) \geq f(\theta_r) \), with strict inequality being the rule. In practice, minorization is carried out piecemeal on a sum of terms defining the objective.

For our particular problem we capitalize on the convexity of the function \(-\ln(s)\). The supporting hyperplane inequality implies the linear minorization

\[
- \sum_{i=1}^{n} \ln(1 + \theta' c_i) \geq - \sum_{i=1}^{n} \frac{1}{1 + \theta'_i c_i} (1 + \theta'_i c_i - 1 - \theta'_i c_i).
\]

On the other hand, Jensen’s inequality gives the minorization

\[
\sum_{i=1}^{n} \ln(1 + \theta' b_i) \geq \sum_{i=1}^{n} \frac{1}{1 + \theta'_i b_i} \ln \left( \frac{1 + \theta'_i b_i}{1} \right) + \sum_{i=1}^{n} \sum_{j} \frac{\theta_{rj} b_{ij}}{1 + \theta'_i b_i} \ln \left( \frac{1 + \theta'_i b_i}{\theta_{rj} b_{ij} \theta_{rj}} \right).
\]

The sum of these two minorizations constitutes the overall minorization \( g(\theta \mid \theta_r) \). The stationary condition \( \nabla g(\theta \mid \theta_r) = 0 \) can be solved to yield the updates

\[
\theta_{r+1,j} = \theta_{rj} \frac{\sum_{i=1}^{n} \frac{b_{ij}}{1 + \theta'_i c_i}}{\sum_{i=1}^{n} \frac{1}{1 + \theta'_i c_i}}.
\]

Note that the update \( \theta_{r+1,j} \) remains nonnegative if \( \theta_{rj} \) is nonnegative and equals 0 if and only if \( \theta_{rj} = 0 \). However, convergence of \( \theta_{rj} \) to 0 is possible. More importantly, the MM updates drive the loglikelihood uphill.

### 6.4 Initialization

Most optimization algorithms benefit from good starting values. The obvious candidate for \( \beta \) is the maximum likelihood estimate delivered by the independence model using \( GLM.jl \). Under the VC framework, we use the MM algorithm to initialize variance components. Under the CS and AR(1) framework, we initialize the variance component \( \sigma^2 \) by the crude estimate from the MM algorithm treating \( \rho = 1 \).

### 7 Statistical Properties

Because the likelihood is a smooth function of the parameters in the quasi-copula model, we expect the maximum likelihood estimates \((\hat{\beta}, \hat{\theta})\) to be consistent and
asymptotically efficient. One can estimate the asymptotic covariance matrix by the inverse of the observed information matrix. The expected information matrix is probably unavailable in closed form. It is straightforward to implement likelihood ratio tests on the mean components \( \beta \). Likelihood ratio testing on the variance components \( \theta \) is complicated by the same nonnegativity constraints implicit in all variance components models.

### 7.1 Compound Symmetric Covariance

Under the Compound Symmetric (CS) parameterization of \( \Gamma_i \), one can test hypotheses involving the correlation parameter \( \rho \). To ensure that the covariance matrix \( \Gamma_i \) is positive semi-definite, we bound \( \rho \in (-\frac{1}{d_i - 1}, 1) \). Additional details on this derivation can be found in the supplemental materials. For example, in the bivariate case, \( d_i = 2 \), \( \rho \in (-1, 1) \) and

\[
\Gamma_i = \sigma^2 \times \begin{bmatrix} \rho & 1 \\ 1 & \rho \end{bmatrix}
\]

We are interested in the hypothesis \( H_0 : \rho = 0 \), which represents an independent univariate generalized linear mixed model with a single variance component proportional to the identity matrix. The additional noise component captures overdispersion.

### 8 Results

#### 8.1 Simulation Studies

To assess estimation accuracy of the quasi-copula model, we first present simulation studies for the Poisson and negative binomial base distributions with log link function, under the VC parameterization of \( \Gamma_i \). We then demonstrate the flexibility of the model in analyzing mixed discrete outcomes under a bivariate model with Poisson and Bernoulli base distributions and canonical link functions. Additional simulation studies with different base distributions under the AR(1), CS and VC parameterizations of \( \Gamma_i \) are included in the supplemental material.

In each simulation scenario, the non-intercept entries of the predictor matrix \( X_i \) are independent standard normal deviates. True regression coefficients \( \beta_{\text{true}} \sim \text{Uniform}(-0.2, 0.2) \). For the negative binomial base, all dispersion parameters are
$r_{\text{true}} = 10$. Each simulation scenario was run on 100 replicates for each sample size $n \in \{100, 1000, 10000\}$ and number of observations $d_i \in \{2, 5, 10, 15, 20, 25\}$ per independent sampling unit.

Under the VC parameterization of $\Gamma_i$, the choice $\Gamma_i, \text{true} = \theta_{\text{true}} \times 1_{d_i}1_{d_i}^T$ allows us to compare to the random intercept GLMM fit using MixedModels.jl. When the random effect term is a scalar, MixedModels.jl uses Gaussian quadrature for parameter estimation. We compare estimates and run-times to the random intercept GLMM fit of MixedModels.jl with 25 Gaussian quadrature points. We conduct simulation studies under two scenarios (simulation I and II). In simulation I, it is assumed that the data are generated by the quasi-copula model with $\theta_{\text{true}} = 0.1$, and in simulation II, it is assumed that the true distribution is the random intercept GLMM with $\theta_{\text{true}} = 0.01, 0.05$.

**Simulation I:** In this scenario, we simulate datasets under the quasi-copula model as outlined in Section 5 and compare MLE fits under the quasi-copula model and GLMM. Figures 1-2 help us assess estimation accuracy and how well the GLMM density approximates the quasi-copula density. As anticipated, the MSE's across all base distributions decrease as sample size increases. For data simulated under the quasi-copula model, quasi-copula mean squared errors (MSE) are generally lower than GLMM MSE's. GLMM estimated variance components are often zero and stay relatively constant across sample sizes. This confirms the fact that the two models are different in how they handle random effects, particularly with larger sampling units ($d_i > 2$).

**Simulation II:** In the second simulation scenario, we generate datasets under the random intercept Poisson GLMM and compare MLE fits delivered by the two models. Figures 3-6 now shed light on how well the quasi-copula density approximates the GLMM density under different magnitudes of the variance components. As expected, MSE's under GLMM analysis are now generally lower than those under quasi-copula analysis. For the Poisson and negative binomial base distributions with $\theta_{\text{true}} = 0.05$, Figures 3-4 indicate biases for the quasi-copula estimates of $(\beta, \theta)$ for larger sampling units ($d_i > 2$) up to sample size $n = 10,000$. However in Figures 5-6, where the variance component $\theta_{\text{true}} = 0.01$ is smaller, we no longer observe biased estimates for $(\beta, \theta)$. 
Figure 1: Simulation I: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the Poisson base distribution with log link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario result includes 100 replicates.
Figure 2: Simulation I: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the negative binomial base distribution with log link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario result includes 100 replicates.
Figure 3: Simulation II: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta = 0.05$ under the Poisson base distribution with log link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario reports involves 100 replicates.
Figure 4: Simulation II: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta = 0.05$ under the negative binomial base distribution with log link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario result includes 100 replicates.
Figure 5: Simulation II: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta = 0.01$ under the Poisson base distribution with log link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario result includes 100 replicates.
Figure 6: Simulation II: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta = 0.01$ under the negative binomial base distribution with log link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario result includes 100 replicates.
8.1.1 Run Times

Run times under simulation I and II are comparable. Table 1 presents average run times and their standard errors in seconds for 100 replicates under simulation II with $\theta_{true} = 0.01$. All computer runs were performed on a standard 2.3 GHz Intel i9 CPU with 8 cores. Runtimes for the quasi-copula model are presented given multi-threading across 8 cores. We note the current version of MixedModels.jl does not allow for multi-threading across multiple cores. Because in contrast to MixedModels.jl the quasi-copula loglikelihoods contain no determinants or matrix inverses, QuasiCopula.jl experiences less pronounced increases in computation time as sample and sampling unit sizes grow. Run times for the quasi-copula model are faster than those of MixedModels.jl for discrete outcomes (Table 1, Supplementary Table 5) and slower for Gaussian distributed outcomes (Supplementary Table 6). This general trend also holds on a per core basis. This discrepancy is hardly surprising since MixedModels.jl takes into account the low-rank structure of the covariance matrix $\Omega_i$ in the random intercept linear mixed model (LMM). This tactic reduces the computational complexity per sample from $O(d_i^3)$ to $O(d_i^2)$. More detailed comparisons appear in the supplement.

For the negative binomial base distribution, MixedModels.jl explicitly warns the user against fitting GLMM’s with unknown dispersion parameter $r$. Our software updates $r$ iteratively by Newton’s method, holding the other parameters ($\beta, \theta$) fixed. Our restriction to MixedModels.jl makes for a fair comparison within the Julia language universe. We also compared our negative binomial fits with those delivered by the three popular R packages for GLMM estimation in Tables 2. On a single dataset with $d_i = 5$, and $n = 10,000$ simulated under simulation II, the lme4 package [Bates et al., 2015] takes an inordinately long time to fit the model. Obtaining confidence intervals takes a significant amount of additional time, and inference of $r$ is impossible. The glmmTMB package [Brooks et al., 2017] allows for inference of $r$ and takes much less time to form confidence intervals than lme4, but it is still significantly slower than quasi-copula fitting. Both lme4 and glmmTMB fit the negative binomial GLMM using Laplace Approximation, while the GLMMadaptive package [Pinheiro and Bates, 1995] uses adaptive Gaussian quadrature. In Tables 2 and 3, we use GLMMadaptive to fit the data with 25 Gaussian quadrature points. GLMMadaptive allows for inference of $r$ and takes no additional time to form confidence intervals, but is still significantly slower than quasi-copula fitting. Run times in seconds for obtaining the estimates and confidence intervals in Table 2 appear in Table 3.
Table 1: Run times and (standard error of run times) in seconds based on 100 replicates under simulation II with Poisson and negative binomial (NB) Base, $\theta_{true} = 0.01$, sampling unit size $d_i$ and sample size $n$.

| Parameter | Truth | QC fit | lme4 fit | glmmTMB fit | GLMMadaptive fit |
|-----------|-------|--------|----------|-------------|------------------|
| $\beta_1$ | 0.036 | 0.033  | 0.032    | 0.033       | 0.032            |
|           | (0.028, 0.037) | (0.022, 0.042) | (0.023, 0.043) | (0.023, 0.042) |
| $\beta_2$ | 0.107 | 0.106  | 0.106    | 0.106       | 0.106            |
|           | (0.101, 0.111) | (0.097, 0.115) | (0.097, 0.115) | (0.097, 0.115) |
| $\beta_3$ | 0.026 | 0.026  | 0.026    | 0.026       | 0.026            |
|           | (0.017, 0.035) | (0.017, 0.035) | (0.017, 0.035) | (0.017, 0.035) |
| $\theta$ | 0.01  | 0.007  | 0.009    | 0.008       | 0.009            |
|           | (0.003, 0.011) | (0.002, 0.015) | (0.002, 0.019) | (0.005, 0.018) |
| $r$       | 10    | 10.002 | 10.147   | 10.101      | 9.996            |
|           | (9.094, 10.910) | (NA, NA) | (8.640, 11.809) | (8.612, 11.602) |

Table 2: MLE’s and (confidence intervals) based on a single replicate under simulation II with negative binomial Base, $\theta_{true} = 0.01$, sampling unit size $d_i = 5$ and sample size $n = 10000$.

Table 3: Run times and (confidence interval run times) in seconds based on a single replicate under simulation II with negative binomial Base, $\theta_{true} = 0.01$, sampling unit size $d_i = 5$ and sample size $n = 10000$. 

| n  | $d_i$ | QC time | lme4 time | glmmTMB time | GLMMadaptive time |
|----|------|---------|-----------|--------------|-------------------|
| 10000 | 5    | 0.021 (0.001) | 0.022 (0.003) | 0.125 (0.008) | 0.037 (0.003) |
| 10000 | 2    | 0.025 (0.001) | 0.045 (0.003) | 0.085 (0.005) | 0.068 (0.004) |
| 10000 | 10   | 0.023 (0.001) | 0.080 (0.004) | 0.105 (0.004) | 0.187 (0.011) |
| 10000 | 15   | 0.024 (0.001) | 0.148 (0.006) | 0.165 (0.004) | 0.282 (0.017) |
| 10000 | 20   | 0.025 (0.001) | 0.186 (0.007) | 0.112 (0.002) | 0.394 (0.017) |
| 10000 | 25   | 0.026 (0.001) | 0.265 (0.009) | 0.119 (0.003) | 0.461 (0.019) |
| 10000 | 5    | 0.030 (0.001) | 0.516 (0.016) | 0.167 (0.004) | 0.857 (0.033) |
| 10000 | 10   | 0.035 (0.001) | 1.011 (0.022) | 0.243 (0.003) | 1.972 (0.050) |
| 10000 | 15   | 0.040 (0.001) | 1.402 (0.030) | 0.303 (0.002) | 2.854 (0.064) |
| 10000 | 20   | 0.042 (0.001) | 1.887 (0.036) | 0.371 (0.002) | 3.722 (0.077) |
| 10000 | 25   | 0.051 (0.001) | 2.531 (0.046) | 0.435 (0.002) | 4.815 (0.089) |
| 10000 | 2    | 0.128 (0.001) | 1.896 (0.032) | 1.169 (0.040) | 3.902 (0.079) |
| 10000 | 5    | 0.154 (0.001) | 4.333 (0.075) | 1.375 (0.020) | 8.598 (0.140) |
| 10000 | 10   | 0.232 (0.002) | 9.545 (0.143) | 2.154 (0.007) | 20.499 (0.303) |
| 10000 | 15   | 0.272 (0.002) | 14.844 (0.249) | 2.78 (0.007) | 29.083 (0.465) |
| 10000 | 20   | 0.336 (0.002) | 21.423 (0.356) | 3.314 (0.007) | 42.952 (0.679) |
| 10000 | 25   | 0.429 (0.003) | 29.324 (0.528) | 4.111 (0.007) | 54.676 (0.861) |
8.2 Bivariate Mixed Outcome Model

Let \( y_i = (y_{i1}, y_{i2})^t \) denote the \( i \)th bivariate mixed discrete outcome from \( n \) bivariate sampling units. For purposes of illustration we assume that \( y_{i1} \) follows a Poisson base distribution and \( y_{i2} \) follows a Bernoulli base distribution under their canonical link functions. Thus,

\[
y_{i1} \sim \text{Poisson} [\mu_{i1}(\beta_1)], \quad \text{where } \log[\mu_{i1}(\beta_1)] = x_i^t \beta_1 \\
y_{i2} \sim \text{Bernoulli} [\mu_{i1}(\beta_2)], \quad \text{where } \logit[\mu_{i2}(\beta_2)] = x_i^t \beta_2.
\]

For each independent realization \( y_i \), we postulate a vector of covariates \( x_i \) and a corresponding vector of fixed effects, both of length \( p \). An intercept is included among the fixed effects. The fixed effects \( \beta = [\beta_1^T, \beta_2] \) for both responses are jointly estimated under the design matrix \( X_i = [x_i^T, 0_p^T, x_i^T] \). Estimation of the variance components \( \theta \) is unchanged. As expected, Figure 7 shows that all MSEs decrease as the sample size \( n \) increases.

8.3 NHANES Data Example

For many repeated measurement problems, a simple random intercept model is sufficient to account for correlations between different responses on the same subject. To illustrate this point and the performance of the quasi-copula model, we now turn to a bivariate example from the NHANES I Epidemiologic Followup Study (NHEFS) dataset (Cohen, 1987). In this example, we group the data by subject ID and jointly model the number of cigarettes smoked per day in 1971 and the number of cigarettes smoked per day in 1982 as a bivariate outcome. For fixed effects, we include an intercept and control for sex, age in 1971, and the average price of tobacco in the state of residence. The average price of tobacco is a time-dependent covariate that is adjusted for inflation using the 2008 U.S. consumer price index (CPI). Participants with missing responses or predictors were excluded from the model cohort. A total of \( n = 1537 \) NHANES I participants constitute the cohort. Table 4 compares the estimates, loglikelihoods and run times in seconds of the random intercept regression model with Poisson, negative binomial, and Bernoulli base distributions under QuasiCopula.jl and MixedModels.jl. For the Bernoulli base distribution, we transformed each count outcome to a binary indicator with value 1 if the number of cigarettes smoked per day is greater than the sample average and value 0 otherwise.

Because overdispersion is a feature of this dataset, the Poisson base distribution represents a case of model misspecification; the negative binomial base distribution
Figure 7: Simulation I: Mean squared errors (MSE) of MLE estimates $\beta$ and $\theta = 0.1$ under the Bivariate Mixed Outcome model with Poisson and Bernoulli base distributions and their canonical link functions. Each scenario reports involves 100 replicates.
Table 4: Random intercept MLEs, loglikelihoods, and run times for the NHEFS data under the quasi-copula (QC) model and GLMM. All \( n = 1537 \) sampling units are of size \( d_i = 2 \).

| Parameter | QC Poisson | GLMM Poisson | QC NB | GLMM NB | QC Bernoulli | GLMM Bernoulli |
|-----------|------------|--------------|-------|---------|--------------|----------------|
| \( \beta_{\text{intercept}} \) | -2.509 | -2.032 | -2.980 | -2.980 | -1.768 | -1.411 |
| \( \beta_{\text{sex}} \) | -0.210 | -0.225 | -0.187 | -0.187 | -0.605 | -0.701 |
| \( \beta_{\text{age}} \) | -0.434 | -0.397 | -0.407 | -0.407 | -2.238 | -1.891 |
| \( \beta_{\text{price}} \) | 0.434 | 0.597 | 0.402 | 0.402 | 2.238 | 1.891 |
| \( \theta \) | 7.080 | 0.458 | 0.0 | 0.0 | 0.666 | 3.461 |
| \( r \) | -1.141 | 1.795 | - | - | - | - |
| loglikelihood | -20690.797 | -15499.537 | -12037.587 | -12047.504 | -1938.712 | -1980.893 |
| Time (seconds) | 0.268 | 0.749 | 0.160 | 0.978 | 0.109 | 1.030 |

is a better choice for analysis. Under the Poisson base distribution, the quasi-copula model inflates the variance component to account for the overdispersion. Under the negative binomial base distribution, both QuasiCopula.jl and MixedModels.jl estimate the variance component to be 0. This suggests that no additional overdispersion exists in the data. The estimates for \( \beta \) under the quasi-copula model with Poisson base are closer to the more realistic estimates under the negative binomial base than those of GLMM. The maximum loglikelihood of the quasi-copula model is lower than that of GLMM for the Poisson base and higher than that of GLMM for the negative binomial and Bernoulli bases. Run times favor the quasi-copula model.

9 Discussion

We propose a new model for analyzing multivariate data based on Tonda’s Gaussian copula approximation. Our quasi-copula model enables the analysis of correlated responses and handles random effects needed in applications such as panel and repeated measures data. The quasi-copula model trades Tonda’s awkward parameter space constraint for a simple normalizing constant. This allows one to engage in full likelihood analysis under a tractable probability density function with no implicit integrations or matrix inverses. The quasi-copula model is relatively easy to fit and friendly to likelihood ratio hypothesis testing. Additionally, it easily extends to accommodate mixtures of different base distributions.

For maximum likelihood estimation, we recommend a combination of two numerical methods. The first is a block ascent algorithm that alternates between updating the mean parameters \( \beta \) by a version of Newton’s method and updating the variance components by a minorization-maximization (MM) algorithm. The second method jointly updates \( \beta \) and the variances components by a standard quasi-Newton algorithm. The MM algorithm converges quickly to a neighborhood of the MLE but then slows. In contrast, the quasi-Newton struggles at first and then converges quickly.
Thus, we start with the block ascent algorithm and then switch to the quasi-Newton algorithm. Both algorithms and their combination are available in our QuasiCopula.jl Julia package.

On balance our numerical tests suggest limitations of the quasi-copula model in handling strongly correlated responses and large sampling units. The presence and size of the normalizing constant $1 + \text{tr}(\Gamma)$ in the quasi-copula density may well be the culprit. When the true distribution follows the random intercept GLMM, the quasi-copula estimates are most accurate for small sampling units. When sampling units are large, the quasi-copula estimates are reasonably accurate for smaller magnitudes of variance components. In actual practice many statisticians simply assume the validity of their underlying statistical model. We sorely need good methods for assessing model appropriateness. In the meanwhile, the quasi-copula model offers another avenue for analysis of correlated data. In our opinion, its speed and versatility make up for it defects. We hope other statisticians will agree and assist in probing its properties and applying it to challenging datasets.

Acknowledgements

We thank Tetsuji Tonda for the inspiration behind our new class of models.

Disclosure statement

The authors have no conflicts of interest.

10 Web Resources

Our software QuasiCopula.jl is freely available to the scientific community through the OpenMendel (Zhou et al., 2020) platform.

Project name: QuasiCopula.jl
Project home page: https://github.com/OpenMendel/QuasiCopula.jl
Supported operating systems: Mac OS, Linux, Windows
Programming language: Julia 1.6, 1.7
License: MIT

All commands needed to reproduce the following results are available at https://github.com/sarah-ji/QuasiCopula-reproducibility
References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1):1–48.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American statistical Association, 88(421):9–25.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R Journal, 9(2):378–400.

Cohen, B. B. (1987). Plan and operation of the NHANES I Epidemiologic Followup Study, 1982-84. Number 22. US Department of Health and Human Services, Public Health Service, National . . . .

Lange, K. (2016). MM Optimization Algorithms. SIAM.

Lange, K., Hunter, D. R., and Yang, I. (2000). Optimization transfer using surrogate objective functions. Journal of Computational and Graphical Statistics, 9(1):1–20.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1):13–22.

Ohtaki, M. (1999). Globally convergent algorithm without derivatives for maximizing a multivariate function. Proceedings of Development of Statistical Theories and their Application for Complex Nonlinear Data.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics, 4(1):12–35.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris, 8:229–231.

Song, P. X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using gaussian copulas. Biometrics, 65(1):60–68.

Tonda, T. (2005). A class of multivariate discrete distributions based on an approximate density in \{GLMM\}. Hiroshima Mathematical Journal, 35(2):327–349.
Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. Journal of the American statistical association, 86(413):79–86.

Zhou, H., Hu, L., Zhou, J., and Lange, K. (2019). Mm algorithms for variance components models. Journal of Computational and Graphical Statistics, 28(2):350–361.

Zhou, H., Sinsheimer, J. S., Bates, D. M., Chu, B. B., German, C. A., Ji, S. S., Keys, K. L., Kim, J., Ko, S., Mosher, G. D., et al. (2020). OpenMendel: a cooperative programming project for statistical genetics. Human Genetics, 139(1):61–71.
11 Supplemental Materials

11.1 Tonda’s Approximation Details

Let \( x \) be a random vector with exponential density \( f(x \mid \nu) = e^{T(x)^t \nu - A(\nu)} \). Note that \( T(x) \) has mean \( \mu(\nu) = \nabla A(\nu) \) and covariance matrix \( d^2 A(\nu) \). Let us shift \( \nu \) by adding a random Gaussian \( z \) with mean \( 0 \) and covariance \( \Sigma \). The new density \( E[e^{T(x)^t (\nu + z) - A(\nu + z)}] \) can be approximated by expanding the integrand to second order around \( z = 0 \) and integrating. This yields

\[
E[e^{T(x)^t (\nu + z) - A(\nu + z)}] \approx E\left(e^{T(x)^t \nu - A(\nu)} \{1 + [T(x) - \nabla A(\nu)][T(x) - \nabla A(\nu)]^t \} \right)
\]

\[
= e^{T(x)^t \nu - A(\nu)} \left(1 + \frac{1}{2} \text{tr}\{[T(x) - \mu(\nu)][T(x) - \mu(\nu)]^t \Sigma\} \right)
\]

\[
= e^{T(x)^t \nu - A(\nu)} \left(1 + \frac{1}{2} \text{tr}\{d^2 A(\nu)\}\right)
\]

where \( W \) is the standardized version \([T(x) - \mu(\nu)]d^2 A(\nu)^{-1/2}\) of the base sufficient statistic \( T(x) \). The condition \( 1 - \frac{1}{2} \text{tr}\{[T(x) - \mu(\nu)]d^2 A(\nu)\} > 0 \) is sufficient but not necessary for the approximate density to be nonnegative. When this condition holds, the approximate density has mass 1. In our quasi-copula density, we drop the offending term \(-\frac{1}{2} \text{tr}\{[T(x) - \mu(\nu)]d^2 A(\nu)\}\), replace \( \sqrt{d^2 A(\nu)} \Sigma \sqrt{d^2 A(\nu)} \) by \( \Gamma \), assume \( T(x) = x \), and normalize.

11.2 Generate Random Deviates

We can construct the \( d \) dimensional multivariate vector, \( y \) from the multivariate density \( g_y(y) \) element wise using conditional densities. We recognize the joint density
can be represented as a product of conditional densities:

\[ g_y(y) = g_{y_1}(y_1) \times g_{y_2|y_1}(y_2|y_1) \times \cdots \times g_{y_d|y_1,\ldots,y_{d-1}}(y_d|y_1, \ldots, y_{d-1}) \]

Thus we can first sample \( y_1 \) from its marginal density \( g_{y_1}(y_1) \), and then sample \( y_2 \) from the conditional density \( g_{y_2|y_1}(y_2|y_1) \). The resulting set is a sample from the joint density \( g_{y_1,y_2}(y_1, y_2) \). Continuing this process for all \( n \) values of the multivariate vector, \( y \), we can sample from its joint density \( g_y(y) \).

11.2.1 Marginal Distribution

For every univariate base distribution, the required probability density functions (PDFs) \( g_y(y) \) are of the same form, where \( c_0, c_1 \) and \( c_2 \) are constants that depend on the parameters of the specified base distribution \( f_y(y) \).

\[
 g_y(y) = cf_y(y)\left[a_0 + a_1(y - \mu) + a_2(y - \mu)^2\right] \quad (2)
 = cf_y(y) \times \left[c_0 + c_1y + c_2y^2\right], \quad (3)
\]

We can re-arrange the PDF to derive the constants \( c_0, c_1, c_2 \) in the marginal PDF \( g_y(y) \) as follows:

\[
 g_y(y) = \left(1 + \frac{1}{2}\text{tr}(\Gamma)\right)^{-1} f_y(y) \left[1 + \frac{\gamma_{11}}{2} \left(y - \frac{\mu}{\sigma}\right)^2 + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj}\right]
 = \left(1 + \frac{1}{2}\text{tr}(\Gamma)\right)^{-1} f_y(y) \left[1 + \frac{\gamma_{11}}{2} \frac{y^2 - 2y\mu + \mu^2}{\sigma^2} + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj}\right]
 = \left(1 + \frac{1}{2}\text{tr}(\Gamma)\right)^{-1} f_y(y) \left[1 + \frac{\gamma_{11}}{2} \frac{y^2}{\sigma^2} + \frac{\gamma_{11}}{2} \frac{-2y\mu}{\sigma^2} + \frac{\gamma_{11}}{2} \frac{\mu^2}{\sigma^2} + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj}\right]
 = \left(1 + \frac{1}{2}\text{tr}(\Gamma)\right)^{-1} f_y(y) \left[1 + \frac{\gamma_{11}}{2} \frac{\mu^2}{\sigma^2} + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj}\right] + \left(\frac{\gamma_{11}}{2} \frac{-2\mu}{\sigma^2}\right) y + \left(\frac{\gamma_{11}}{2} \frac{1}{\sigma^2}\right) y^2
 = c \times f_y(y) \left[c_0 + c_1y + c_2y^2\right],
\]

\[ c = \left[1 + \frac{1}{2}\text{tr}(\Gamma)\right]^{-1}, \text{ for all base distributions } f_y(y). \]
\[ c_0 = \left( 1 + \frac{\gamma_1}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right), \]
\[ c_1 = \left( \frac{\gamma_1}{2} \left( \frac{-2\mu}{\sigma^2} \right) \right), \]
\[ c_2 = \left( \frac{\gamma_1}{2} \left( \frac{1}{\sigma^2} \right) \right). \]

Thus, the required marginal Cumulative Distribution Function (CDF) \( G_y(x) \) takes the following form. We will derive the CDF by finding the appropriate scaled cumulative distributions of the three terms.

\[
G_y(x) = \int_{0}^{\infty} g_y(y) \, dy
= c \int_{-\infty}^{x} f(y) [c_0 + c_1 y + c_2 y^2] \, dy
= c \times c_0 \int_{-\infty}^{x} f_y(y) \, dy
+ c \times c_1 \int_{-\infty}^{x} y f_y(y) \, dy
+ c \times c_2 \int_{-\infty}^{x} y^2 f_y(y) \, dy
= \text{term1} + \text{term2} + \text{term3}
\]

The first term is a scalar multiple of the base distribution CDF, \( F_y(y) \), and \( d_1, d_2 \) are normalizing constants for random variables \( v_1, v_2 \) from named distributions \( f_{v_1}(v_1), f_{v_2}(v_2) \) with CDFs \( F_{v_1}(x) \) and \( F_{v_2}(x) \) in terms 2 and 3, respectively.

- \text{term1} = c \times c_0 \int_{-\infty}^{x} f_y(y) \, dy = c \times (c_0) \times F_y(x),
- \text{term2} = c \times c_1 \int_{-\infty}^{x} y \cdot f_y(y) \, dy = c \times c_1 \times d_1 \times \int_{-\infty}^{x} f_{v_1}(y) \, dy = c \times (c_1) \times d_1 \times F_{v_1}(x)
- \text{term3} = c \times c_2 \int_{-\infty}^{x} y^2 \cdot f_y(y) \, dy = c \times c_2 \times d_2 \times \int_{-\infty}^{x} f_{v_2}(y) \, dy = c \times (c_2) \times d_2 \times F_{v_2}(x)

For every base distribution, to satisfy properties of a proper distribution function we require
\[
c \times [c_0 + c_1 \times d_1 + c_2 \times d_2] = 1.
\]
11.2.2 Conditional Distribution

Let $y_{[i-1]}$ indicate elements $y_1, ..., y_{i-1}, \forall i \in [1, d]$. Then the conditional density of $y_i$ given the previous components $y_{[i-1]}$ is:

$$g_{y_i|y_{[i-1]}}(y_i|y_{[i-1]}) = d_{[i-1]}^{-1} f_i(y_i) \left[ d_{[i-1]} + r_i \sum_{j=1}^{i-1} r_j \gamma_{ij} + \frac{\gamma_{ii}}{2} (r_i^2 - 1) \right]$$

$$= d_{[i-1]}^{-1} f_i(y_i) \left[ d_{[i-1]} + \left( \frac{y_i - \mu_i}{\sigma_i} \right) \sum_{j=1}^{i-1} r_j \gamma_{ij} + \frac{\gamma_{ii}}{2} (r_i^2 - 1) \right]$$

$$= d_{[i-1]}^{-1} f_i(y_i) \left[ \left( d_{[i-1]}^{-1} - \frac{\gamma_{ii}}{2} \right) + \frac{\sum_{j=1}^{i-1} r_j \gamma_{ij}}{\sigma_i} y_i - \mu_i \left( \frac{\sum_{j=1}^{i-1} r_j \gamma_{ij}}{\sigma_i} \right) + \frac{\gamma_{ii}}{2} \frac{(y_i - \mu_i)^2}{\sigma_i^2} \right]$$

$$= d_{[i-1]}^{-1} f_i(y_i) \left[ \left( d_{[i-1]}^{-1} - \frac{\gamma_{ii}}{2} \right) - \mu_i \left( \frac{\sum_{j=1}^{i-1} r_j \gamma_{ij}}{\sigma_i} \right) + \frac{\gamma_{ii}}{2} \frac{\mu_i^2}{\sigma_i^2} \right]$$

$$+ \left( \frac{\gamma_{ii}}{2} \left( \frac{1}{\sigma_i^2} \right) \right) y_i^2$$

$$= cf_i(y_i) \left[ c_0 + c_1 y_i + c_2 y_i^2 \right]$$

where $d_{[i-1]} = 1 + \frac{1}{2} \tau_{[i-1]} \Gamma_{[i-1]} r_{[i-1]} + \frac{1}{2} \sum_{j=1}^{d} \gamma_{jj}$.

- $c = d_{[i-1]}^{-1}$
- $c_0 = \left( d_{[i-1]}^{-1} - \frac{\mu_i}{2} - \mu_i \left( \frac{\sum_{j=1}^{i-1} r_j \gamma_{ij}}{\sigma_i} \right) + \frac{\gamma_{ii}}{2} \frac{\mu_i^2}{\sigma_i^2} \right)$
- $c_1 = \left( \frac{\sum_{j=1}^{i-1} r_j \gamma_{ij}}{\sigma_i} + \frac{\mu_i}{2} \left( \frac{-2 \mu_i}{\sigma_i^2} \right) \right)$
- $c_2 = \left( \frac{\gamma_{ii}}{2} \left( \frac{1}{\sigma_i^2} \right) \right)$

We can construct each conditional density given the previously sampled elements as a combination of three constants, just as in the marginal density.
11.2.3 Continuous Outcomes

When marginal densities \( f(y) \) are continuous, each stage of sampling is probably best performed by inverse transform sampling.

**Gamma distribution**  This note considers the special case of Gamma base in the copula framework outlined in Ken’s notes, where \( f_1(y) \sim \Gamma(\alpha, \theta) \). We will simulate \( y \) directly from its marginal density, \( g_y(y) \), which can also be represented as a mixture distribution.

\[
y \sim \Gamma(\alpha, \theta); 
\]  \( f_y(y) = \frac{1}{\Gamma(\alpha)\theta^\alpha} y^{\alpha-1} e^{-\frac{y}{\theta}} \)

- \( \mu = E[y] = \alpha \theta \), and \( \sigma^2 = Var(y) = \alpha \theta^2 \).

\[
g_y(y) = \left[1 + \frac{1}{2} \text{tr}(\Gamma)^{-1} \left[ \frac{1}{\Gamma(\alpha)\theta^\alpha} y^{\alpha-1} e^{-\frac{y}{\theta}} \right] \left(1 + \frac{\gamma_{11}}{2} \left( \frac{y - \mu}{\sigma^2} \right)^2 \right) + \frac{1}{2} \sum_{j=2}^d \gamma_{jj} \right] 
\]

\[
= \left[1 + \frac{1}{2} \text{tr}(\Gamma)^{-1} \left[ \frac{1}{\Gamma(\alpha)\theta^\alpha} y^{\alpha-1} e^{-\frac{y}{\theta}} \right] \left(1 + \frac{1}{2} \sum_{j=2}^d \gamma_{jj} \right) + \frac{\gamma_{11}}{2} \left( \frac{y^2 - 2y\mu + \mu^2}{\sigma^2} \right) \right] 
\]

\[
= c \times f_y(y) \left[ (c_0) + (c_1) y + (c_2) y^2 \right] 
\]

- Here \( c_0 = \left(1 + \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^d \gamma_{jj} \right) \)

\[
= \left(1 + \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^d \gamma_{jj} \right) 
\]

\[
= \left(1 + \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^d \gamma_{jj} \right) 
\]

\[
= \left(1 + \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^d \gamma_{jj} \right) 
\]

- \( c_1 = \left( \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) \right) \)

\[
= \left( \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) \right) 
\]

- \( c_2 = \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\sigma^2} \right) \right) \)

\[
= \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\sigma^2} \right) \right) 
\]

We will use the given information here: \( y \sim \Gamma(\alpha, \theta) \), where \( F_Y(x) = P(y \leq x) \) to derive the CDF \( G_Y(x) \) term by term.

\[
\text{term1} = c \times (c_0) \int_0^x f_y(y) \, dy 
\]

\[
= c \times (c_0) \times F_Y(Y = x) 
\]

\[
= c \times \left(1 + \frac{\gamma_{11}}{2} \left( \alpha \right) + \frac{1}{2} \sum_{j=2}^m \gamma_{jj} \right) \times F_Y(Y = x) 
\]

34
Define a new random variable \( v_1 \sim \Gamma(\alpha + 1, \theta) \), where \( F_{v_1}(x) = P(v_1 \leq x) \).

**term 2**

\[
\begin{align*}
term_2 & = c \times (c_1) \times \int_0^x y f_y(y) \, dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( -\frac{2}{\theta} \right) \right) \times \int_0^x y f_y(y) \, dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( -\frac{2}{\theta} \right) \right) \times \theta \int_0^x \frac{y}{\theta} f_y(y) \, dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( -\frac{2}{\theta} \right) \right) \times \theta \times \frac{1}{\Gamma(\alpha + 1)} \int_0^x \left[ \frac{1}{\Gamma(\alpha + 1)} y^{(\alpha + 1) - 1} e^{-\frac{y}{\theta}} \right] dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( -\frac{2}{\theta} \right) \right) \times \theta \times \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)} \times F_{v_1}(x)
\end{align*}
\]

Define another random variable \( v_2 \sim \Gamma(\alpha + 2, \theta) \); with CDF \( F_{v_2}(x) = P(v_2 \leq x) \).

**term 3**

\[
\begin{align*}
term_3 & = c \times (c_2) \times \int_0^x y^2 f_y(y) \, dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\alpha \theta^2} \right) \right) \times \int_0^x y^2 f_y(y) \, dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\alpha \theta^2} \right) \right) \times \theta^2 \int_0^x \frac{y^2}{\theta^2} f_y(y) \, dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\alpha \theta^2} \right) \right) \times \theta^2 \times \int_0^x \left[ \frac{1}{\Gamma(\alpha + 2)} y^{(\alpha + 2) - 1} e^{-\frac{y}{\theta}} \right] dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\alpha \theta^2} \right) \right) \times \theta^2 \times \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)} \times \int_0^x \left[ \frac{1}{\Gamma(\alpha + 2) \theta^{\alpha + 2}} y^{(\alpha + 2) - 1} e^{-\frac{y}{\theta}} \right] dy \\
& = c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\alpha \theta^2} \right) \right) \times \theta^2 \times \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)} \times F_{v_2}(x)
\end{align*}
\]
Exponential distribution  Next, we consider when \( f_y(y) \sim \text{Exponential}\left(\frac{1}{\theta}\right) \). To find the appropriate CDF function under this exponential base, we make note of the relationship between the exponential and gamma densities.

\[
y \sim \text{Exponential}\left(\frac{1}{\theta}\right) \iff y \sim \Gamma(\alpha = 1, \theta \geq 0);
\]

\[
f_y(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}} = \frac{1}{\Gamma(1)\theta^1} y^{1-1} e^{-\frac{y}{\theta}}; y, \theta \geq 0
\]

• \( \mu = E[y] = \theta \), and \( \sigma^2 = \text{Var}(y) = \theta^2 \).

\[
g_y(y) = \left[1 + \frac{1}{2} \text{tr}(\Gamma)^{-1} \left[\frac{1}{\Gamma(1)\theta^1} y^{1-1} e^{-\frac{y}{\theta}}\right]\right] \left(1 + \frac{\gamma_{11}}{2} \left[ \frac{(y - \mu)^2}{\sigma^2} \right] + \frac{1}{2} \sum_{j=2}^d \gamma_{jj}\right)
\]

\[
= \left[1 + \frac{1}{2} \text{tr}(\Gamma)^{-1} \left[\frac{1}{\Gamma(1)\theta^1} y^{1-1} e^{-\frac{y}{\theta}}\right]\right] \left(1 + \frac{1}{2} \sum_{j=2}^d \gamma_{jj}\right) + \gamma_{11} \left[ \frac{y^2 - 2y\mu + \mu^2}{\sigma^2} \right]
\]

\[
= c \times f_y(y) \left[ \left(c_0\right) + \left(c_1\right) y + \left(c_2\right) y^2 \right]
\]

• Here \( c_0 = \left[1 + \frac{\gamma_{11}}{2} \left(\frac{\mu^2}{\sigma^2}\right) + \frac{1}{2} \sum_{j=2}^d \gamma_{jj}\right] = \left[1 + \frac{\gamma_{11}}{2} + \frac{1}{2} \sum_{j=2}^d \gamma_{jj}\right]
\]

• \( c_1 = \frac{\gamma_{11}}{2} \left(\frac{-2\mu}{\sigma^2}\right) = \frac{\gamma_{11}}{2} \left(\frac{-2}{\sigma}\right)\)

• \( c_2 = \frac{\gamma_{11}}{2} \left(\frac{1}{\sigma^2}\right) = \frac{\gamma_{11}}{2} \left(\frac{1}{\theta^2}\right)\)

\( y \sim \text{Exponential}\left(\frac{1}{\theta}\right) = \Gamma(\alpha = 1, \theta) \), with CDF \( F_Y(Y = x) = P(Y \leq x) \)

\[
\text{term}1 = c \times (c_0) \int_0^x f_y(y)dy = c \times (c_0) \times F_y(x) = c \times \left[1 + \frac{\gamma_{11}}{2} + \frac{1}{2} \sum_{j=2}^d \gamma_{jj}\right] \times F_y(x)
\]

36
Define a new random variable \( v_1 \sim \Gamma(\alpha + 1, \theta) = \Gamma(2, \theta) \), with CDF \( F_{v_1}(x) = P(v_1 \leq x) \).

\[
\text{term 2} = c \times (c_1) \times \int_0^x y f_y(y) \, dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{-2}{\theta} \right) \right) \times \int_0^x y f_y(y) \, dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{-2}{\theta} \right) \right) \times \int_0^x y f_y(y) \, dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{-2}{\theta} \right) \right) \times \int_0^x \left[ \frac{1}{\Gamma(1)\theta^2} y^{(1+1)-1} e^{-y} \right] dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{-2}{\theta} \right) \right) \times \frac{\theta^2}{\theta} \times \frac{\Gamma(2)}{\Gamma(1)} \times \int_0^x \left[ \frac{1}{\Gamma(2)\theta^2} y^{(2)-1} e^{-y} \right] dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{-2}{\theta} \right) \right) \times \frac{\Gamma(2)}{\Gamma(1)} \times F_{v_1}(x)
\]

Define another random variable \( v_2 \sim \Gamma(1 + 2, \theta) = \Gamma(3, \theta) \); with CDF \( F_{v_2}(x) = P(v_2 \leq x) \).

\[
\text{term 3} = c \times (c_2) \times \int_0^x y^2 f_y(y) \, dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\theta^2} \right) \right) \times \int_0^x y^2 f_y(y) \, dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\theta^2} \right) \right) \times \int_0^x \left[ \frac{1}{\Gamma(1)\theta^2} y^{(1+2)-1} e^{-y} \right] dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\theta^2} \right) \right) \times \frac{\theta^3}{\theta} \times \frac{\Gamma(3)}{\Gamma(1)} \times \int_0^x \left[ \frac{1}{\Gamma(3)\theta^3} y^{(3)-1} e^{-y} \right] dy
\]

\[
= c \times \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\theta^2} \right) \right) \times \frac{\Gamma(3)}{\Gamma(1)} \times F_{v_2}(x)
\]

**Beta distribution**  Next, we consider when \( f_y(y) \sim \text{Beta}(\alpha, \beta) \).

\[
y \sim f(y; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad y \in [0, 1]
\]
\[ \mu = E[y] = \frac{\alpha}{\alpha + \beta}, \text{ and } \sigma^2 = Var(y) = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} \]

\[ g_y(y) = \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ \Gamma(\alpha + \beta) \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} y^{\alpha-1} (1-y)^{\beta-1} \right] \left( 1 + \frac{\gamma_{11}}{2} \frac{(y - \mu)^2}{\sigma^2} + \frac{1}{2} \sum_{j=2}^{m} \gamma_{jj} \right) \]

\[ = \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ \Gamma(\alpha + \beta) \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} y^{\alpha-1} (1-y)^{\beta-1} \right] \left( 1 + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) + \frac{\gamma_{11}}{2} \left[ \frac{y^2 - 2y\mu + \mu^2}{\sigma^2} \right] \]

\[ = c \times f_y(y) \left[ (c_0) + \left( c_1 \right) y + \left( c_2 \right) y^2 \right] \]

Here \( c_0 = \left( 1 + \frac{\gamma_{11}}{2} \frac{\mu^2}{\sigma^2} + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) = \left( 1 + \frac{\gamma_{11}}{2} \frac{\alpha (\alpha + \beta + 1)}{\beta} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \)

\( c_1 = \frac{\gamma_{11}}{2} \left( -\frac{2\mu}{\sigma^2} \right) = \frac{\gamma_{11}}{2} \left( -\frac{2(\alpha + \beta)(\alpha + \beta + 1)}{\beta} \right) \)

\( c_2 = \frac{\gamma_{11}}{2} \left( \frac{1}{\sigma^2} \right) = \frac{\gamma_{11}}{2} \left( \frac{(\alpha + \beta)^2 (\alpha + \beta + 1)}{\alpha \beta} \right) \)

\( y \sim \text{Beta}(\alpha, \beta) \) with CDF \( F_Y(Y = x) = P(Y \leq x) \)

\[ \text{term1} = c \times (c_0) \int_{0}^{x} f_y(y)dy \]

\[ = c \times (c_0) \times F_y(x) \]

\[ = c \times \left( 1 + \frac{\gamma_{11}}{2} \left( \frac{\alpha (\alpha + \beta + 1)}{\beta} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) \times F_y(x) \]

Define a new random variable \( v_1 \sim \text{Beta}(\alpha + 1, \beta) \) with CDF \( F_{v_1}(x) = P(v_1 \leq x) \).

\[ \text{term2} = c \times (c_1) \times \int_{0}^{x} y f_y(y)dy \]

\[ = c \times \frac{\gamma_{11}}{2} \left( -\frac{2(\alpha + \beta)(\alpha + \beta + 1)}{\beta} \right) \times \int_{0}^{x} y f_y(y)dy \]

\[ = c \times \frac{\gamma_{11}}{2} \left( -\frac{2(\alpha + \beta)(\alpha + \beta + 1)}{\beta} \right) \times \int_{0}^{x} \left[ \Gamma(\alpha + \beta) \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} y^{(\alpha+1)-1} (1-y)^{\beta-1} \right]dy \]

\[ = c \times \frac{\gamma_{11}}{2} \left( -\frac{2(\alpha + \beta)(\alpha + \beta + 1)}{\beta} \right) \times \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + \beta + 1)} \times \int_{0}^{x} f_{v_1}(y)dy \]

\[ = c \times \frac{\gamma_{11}}{2} \left( -\frac{2(\alpha + \beta)(\alpha + \beta + 1)}{\beta} \right) \times \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha + \beta + 1)} \times F_{v_1}(x) \]

38
Define another random variable $v_2 \sim \text{Beta}(\alpha + 2, \beta)$ with CDF $F_{v_2}(x) = P(v_2 \leq x)$.

**term 3**

\[
\begin{align*}
term 3 &= c \times (c_2) \times \int_0^x y^2 f_y(y) \, dy \\
&= c \times \gamma_{11} \left( \frac{(\alpha + \beta)^2(\alpha + \beta + 1)}{\alpha \beta} \right) \times \int_0^x y^2 f_y(y) \, dy \\
&= c \times \gamma_{11} \left( \frac{(\alpha + \beta)^2(\alpha + \beta + 1)}{\alpha \beta} \right) \times \int_0^x \left[ \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} y^{(\alpha + 2) - 1} (1 - y)^{\beta - 1} \right] dy \\
&= c \times \gamma_{11} \left( \frac{2(\alpha + \beta)(\alpha + \beta + 2)}{\beta} \right) \times \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha + 2 + \beta)} \times \int_0^x f_{v_2}(y) dy \\
&= c \times \gamma_{11} \left( \frac{(\alpha + \beta)^2(\alpha + \beta + 2)}{\alpha \beta} \right) \times \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha + 2 + \beta)} \times F_{v_2}(x)
\end{align*}
\]

### 11.2.4 Discrete Outcomes

When the densities $f_i(y_i)$ are discrete, it may be necessary to compute infinite sums involving these probabilities. For example, such sums occur naturally in numerical algorithms developed for Poisson, Geometric, negative binomial variate generation. From a practical standpoint, it is necessary to truncate these infinite sums after a finite number of terms.

Consider any random variable $Z$ with nonnegative integer values, discrete density $p_i = \Pr(Z = i)$, and mean $\nu$. The inverse method of random sampling reduces to a sequence of comparisons. We partition the interval $[0, 1]$ into subintervals with the $i$th subinterval of length $p_i$. To sample $Z$, we draw a uniform random deviate $U$ from $[0, 1]$ and return the deviate $j$ determined by the conditions $\sum_{i=1}^{j-1} p_i \leq U < \sum_{i=1}^j p_i$. There is no need to invoke the distribution of $Z$. The process is most efficient when the largest $p_i$ occur first. This suggests that we let $k$ denote the least integer $\lfloor \nu \rfloor$ and rearrange the probabilities in the order $p_k, p_{k+1}, p_{k-1}, p_{k+2}, p_{k-2}, \ldots$ This tactic is apt put most of the probability mass first and render sampling efficient.

**Poisson Distribution** A Poisson distribution describes the number of independent events occurring within a unit time interval, given the average rate of occurrence $\theta$.

\[
y \sim \text{Poisson}(\theta); f_y(y) = \frac{\theta^y e^{-\theta y}}{y!}, y = 0, 1, 2, 3, \ldots
\]


\begin{itemize}
\item $\mu = E[y] = \theta = \sigma^2 = Var(y)$
\end{itemize}
\begin{align*}
g_y(y) &= \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ \frac{\theta^y e^{-\theta y}}{y!} \right] \left( 1 + \frac{\gamma_{11}}{2} \left[ \frac{(y - \mu)^2}{\sigma^2} \right] + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) \\
&= \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ \frac{\theta^y e^{-\theta y}}{y!} \right] \left( 1 + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) + \frac{\gamma_{11}}{2} \left[ \frac{y^2 - 2y\mu + \mu^2}{\sigma^2} \right] \\
&= c \times f_y(y) \left[ (c_0) + (c_1) y + (c_2) y^2 \right]
\end{align*}

\begin{itemize}
\item Here $c_0 = \left( 1 + \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^{m} \gamma_{jj} \right) = \left( 1 + \frac{\gamma_{11}}{2} \left( 1 \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right)$
\item $c_1 = \left( \frac{\gamma_{11}}{2} \left( \frac{-2\mu}{\sigma^2} \right) \right) = \left( \frac{\gamma_{11}}{2} \left( -2 \right) \right)$
\item $c_2 = \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\sigma^2} \right) \right) = \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\sigma} \right) \right)$
\end{itemize}

**Binomial Distribution** A Binomial distribution characterizes the number of successes in a sequence of independent trials. It has two parameters: ‘$n$’, the number of trials, and ‘$p$’, the probability of success in an individual trial, with the distribution:

\[ y \sim \text{Binomial}(N, p); f_y(y) = \binom{N}{y} p^y (1 - p)^{N-y}, y = 0, 1, 2, ..., N \]

\begin{itemize}
\item $\mu = E[y] = Np; \sigma^2 = Var(y) = Np(1 - p)$
\end{itemize}
\begin{align*}
g_y(y) &= \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ \binom{N}{y} p^y (1 - p)^{N-y} \right] \left( 1 + \frac{\gamma_{11}}{2} \left[ \frac{(y - \mu)^2}{\sigma^2} \right] + \frac{1}{2} \sum_{j=2}^{m} \gamma_{jj} \right) \\
&= \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ \binom{N}{y} p^y (1 - p)^{N-y} \right] \left( 1 + \frac{1}{2} \sum_{j=2}^{m} \gamma_{jj} \right) + \frac{\gamma_{11}}{2} \left[ \frac{y^2 - 2y\mu + \mu^2}{\sigma^2} \right] \\
&= c \times f_y(y) \left[ (c_0) + (c_1) y + (c_2) y^2 \right]
\end{align*}
Here \( c_0 = \left( 1 + \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) = \left( 1 + \frac{\gamma_{11}}{2} \left( \frac{(Np)^2}{Np(1-p)} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) \)

\( c_1 = \left( \frac{\gamma_{11}}{2} \left( -\frac{2\mu}{\sigma^2} \right) \right) = \left( \frac{\gamma_{11}}{2} \left( \frac{-2Np}{Np(1-p)} \right) \right) \)

\( c_2 = \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\sigma^2} \right) \right) = \left( \frac{\gamma_{11}}{2} \left( \frac{1}{Np(1-p)} \right) \right) \)

**Geometric Distribution**  A Geometric distribution characterizes the number of failures before the first success in a sequence of independent Bernoulli trials with success rate ‘\( p \)’.

\[ y \sim \text{Geometric}(p); \quad f_y(y) = (1-p)^y p, \quad y = 0, 1, 2, ... \]

\( \mu = E[y] = \frac{1}{p}; \quad \sigma^2 = Var(y) = \frac{1-p}{p^2} \)

\[ g_y(y) = \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ (1-p)^y p \right] \left( 1 + \frac{\gamma_{11}}{2} \left[ \frac{(y - \mu)^2}{\sigma^2} \right] + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) \]

\[ = \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ (1-p)^y p \right] \left( 1 + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) + \frac{\gamma_{11}}{2} \left[ \frac{y^2 - 2y\mu + \mu^2}{\sigma^2} \right] \]

\[ = c \times f_y(y) \left[ (c_0) + (c_1) y + (c_2) y^2 \right] \]

Here \( c_0 = \left( 1 + \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) = \left( 1 + \frac{\gamma_{11}}{2} \left( \frac{1}{p^2} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) \)

\( c_1 = \left( \frac{\gamma_{11}}{2} \left( -\frac{2\mu}{\sigma^2} \right) \right) = \left( \frac{\gamma_{11}}{2} \left( \frac{-\gamma_{11}}{p^2} \right) \right) \)

\( c_2 = \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\sigma^2} \right) \right) = \left( \frac{\gamma_{11}}{2} \left( \frac{p^2}{1-p} \right) \right) \)
**Negative Binomial Distribution**  A negative binomial distribution describes the number of failures before the ‘r’th success in a sequence of independent Bernoulli trials. It is parameterized by ‘r’, the number of successes, and ‘p’, the probability of success in an individual trial.

\[ y \sim \text{Negative Binomial}(r, p); f_y(y) = \binom{y + r - 1}{y} p^r (1 - p)^y, y = 0, 1, 2, ... \]

- \( \mu = E[y] = \frac{pr}{1 - p}; \sigma^2 = \text{Var}(y) = \frac{pr}{(1 - p)^2} \)

\[
g_y(y) = \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ \binom{y + r - 1}{y} p^r (1 - p)^y \right] \left( 1 + \frac{\gamma_{11}}{2} \frac{(y - \mu)^2}{\sigma^2} + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right)
\]

\[
= \left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left[ \binom{y + r - 1}{y} p^r (1 - p)^y \right] \left( 1 + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) + \frac{\gamma_{11}}{2} \left[ \frac{y^2 - 2y\mu + \mu^2}{\sigma^2} \right]
\]

- Here \( c_0 = \left( 1 + \frac{\gamma_{11}}{2} \left( \frac{\mu^2}{\sigma^2} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) = \left( 1 + \frac{\gamma_{11}}{2} \left( \frac{pr}{(1 - p)^2} \right) + \frac{1}{2} \sum_{j=2}^{d} \gamma_{jj} \right) \)

- \( c_1 = \left( \frac{\gamma_{11}}{2} \left( \frac{-2\mu}{\sigma^2} \right) \right) = \left( \frac{\gamma_{11}}{2} \left( \frac{-2pr}{(1 - p)^2} \frac{1}{pr} \right) \right) \)

- \( c_2 = \left( \frac{\gamma_{11}}{2} \left( \frac{1}{\sigma^2} \right) \right) = \left( \frac{\gamma_{11}}{2} \left( \frac{(1 - p)^2}{pr} \right) \right) \)
11.3 Parameter Estimation:

We extend the Gaussian Base Model to accommodate densities in exponential family of distributions under the generalized linear model (GLM) framework. In this note, we pay close attention to the density-specific quantities which facilitate parameter estimation, and illustrate using the Poisson and Bernoulli density.

11.3.1 Fisher Scoring to Estimate Beta

\[
\mathcal{L}(\beta) = - \sum_{i=1}^{n} \ln \left[ 1 + \frac{1}{2} \text{tr}(\Gamma_i) \right] + \sum_{i=1}^{n} \ln \left\{ 1 + \frac{1}{2} r_i(\beta)^{T} \Gamma_i r_i(\beta) \right\} + \sum_{i=1}^{n} \sum_{j=1}^{n} \ln f_{ij}(y_{ij} | \beta)
\]

For each distribution, the objective function is the loglikelihood (1), and can be viewed as three separate pieces. The last term of the loglikelihood is specific to the hypothesized density, and has first derivative, \( \sum_{i=1}^{n} \sum_{j} \nabla \ln f_{ij}(y_{ij} | \beta) \), and second derivative \( \nabla^{2} L_n(\beta) \), that generalize to the exponential family of distributions.

The score (gradient of the loglikelihood) with respect to \( \beta \) is:

\[
\nabla L_n(\beta) = \sum_{i=1}^{n} \sum_{j} \nabla \ln f_{ij}(y_{ij} | \beta) + \sum_{i=1}^{n} \frac{\nabla r_i(\beta)^{T} \Gamma_i r_i(\beta)}{1 + \frac{1}{2} r_i(\beta)^{T} \Gamma_i r_i(\beta)},
\]

The first term in the gradient, \( \sum_{i=1}^{n} \sum_{j} \nabla \ln f_{ij}(y_{ij} | \beta) \), corresponds to the first derivative of the piece of the loglikelihood, specific to the hypothesized density. We can write this first term as a function of \( W_{1i} \), a diagonal matrix of "working weights".

\[
\sum_{i=1}^{n} \sum_{j} \nabla \ln f_{ij}(y_{ij} | \beta) = \sum_{i=1}^{n} \sum_{j=1}^{d_{i}} \frac{(y_{ij} - \mu_{ij})\mu'_{ij}(\eta_{ij})}{\sigma_{ij}^{2}} \mu_{ij} = \sum_{i=1}^{n} X_{i}^{T} W_{1i}(Y_{i} - \mu_{i})
\]

\[
W_{1i} = \text{Diagonal} \left( \frac{g'(X_{i}^{T} \beta)}{\text{var}(Y_{i} | \mu_{i})} \right) = \begin{pmatrix}
\mu'_{1}(\eta_{1}) & 0 & \cdots & 0 \\
0 & \mu'_{2}(\eta_{2}) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \mu'_{ni}(\eta_{ni}) / \sigma_{ni}^{2}
\end{pmatrix}
\]

Instead of using the exact Hessian, we will use the expected Fisher Information to get an approximation of the Hessian, which is clearly negative semi-definite.

\[
- \sum_{i=1}^{n} X_{i}^{T} W_{2i} X_{i} - \sum_{i=1}^{n} \frac{[\nabla r_i(\beta)^{T} \Gamma_i r_i(\beta)] [\nabla r_i(\beta)^{T} \Gamma_i r_i(\beta)]}{1 + \frac{1}{2} r_i(\beta)^{T} \Gamma_i r_i(\beta)}^{t}
\]
Specifically, we approximate the second derivative of the piece of the loglikelihood particular to the hypothesized density, $\nabla^2 L_n(\beta)$. Using the Expected Fisher Information Matrix, we present this term as a function of another diagonal weight matrix, $W_{2i}$.

$$\nabla^2 L_n(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ \frac{[\mu'_{ij}(\eta_{ij})]^2}{\sigma^2_{ij}} \right] x_{ij}x_{ij}^T - \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{(y_{ij} - \mu_{ij})}{\sigma^2_{ij}} \right) x_{ij}x_{ij}^T$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{(y_{ij} - \mu_{ij})}{\sigma^2_{ij}} \right) \frac{[\mu'(\eta_{ij})]^2}{\sigma^2_{ij}} (d\sigma^2_{ij}/d\mu_{ij}) x_{ij}x_{ij}^T$$

$$FIM_n(\beta) = E[-\nabla^2 L_n(\beta)] = -\sum_{i=1}^{n} \sum_{j=1}^{d_i} \left[ \frac{[\mu'_{ij}(\eta_{ij})]^2}{\sigma^2_{ij}} \right] x_{ij}x_{ij}^T = -\sum_{i=1}^{n} X_i^T W_{2i} X_i.$$

$$W_{2i} = \text{Diagonal} \left( \frac{g'(X_i^T\beta)^2}{\text{var}(Y_i|\mu_i)} \right) = \begin{pmatrix} \frac{[\mu'_{1i}(\eta_{1i})]^2}{\sigma^2_{1i}} & 0 & \cdots & 0 \\ 0 & \frac{[\mu'_{2i}(\eta_{2i})]^2}{\sigma^2_{2i}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{[\mu'_{ini}(\eta_{ini})]^2}{\sigma^2_{ini}} \end{pmatrix}$$

The score and approximate Hessian provide the ingredients for a kind of scoring algorithm for improving $\beta$ in our model. For each Newton update of the fixed effect parameter, $\beta$, in addition to updating the residual vector, $r_i(\beta)$, we must also update these weight matrices, $W_{1i}$ and $W_{2i}$, in the update of the Score and Hessian. We can find these quantities easily by making the appropriate calls to the GLM package, GLM.jl.

Let $y_{ij}$ represent the $j^{th}$ outcome for person $i$, hypothesized to come from a non-normal density in the exponential family of distributions, $f_{ij}(y_{ij} | \beta)$. For each hypothesized density under the GLM framework, we have mean parameter $\mu_{ij}(\beta) = g^{-1}(\eta_{ij}(\beta)) = g^{-1}(x_{ij}\beta)$, and variance parameter $\sigma^2_{ij}(\beta)$. Using these quantities, we define $r_{ij}(\beta), j \in [1, d_i]$ as the $j^{th}$ entry in the standardized residual vector for observation or group $i$.

$$r_{ij}(\beta) = \sqrt{\tau}(y_{ij} - \mu_{ij}(\beta)) = \frac{(y_{ij} - \mu_{ij}(\beta))}{\sqrt{\sigma^2_{ij}(\beta)}} \in \mathbb{R} \quad (7)$$

Let $\nabla r_i(\beta) \in \mathbb{R}^{d_i \times p}$ denote the matrix of differentials of all $d_i$ observations for the $i^{th}$ individuals standardized residual vector $r_i(\beta)$. This quantity is important for
our score and hessian computation, which really helps the optimization algorithm to speed up convergence.

\[ \nabla r_i(\beta)^t = (\nabla r_{i1}(\beta) \quad \nabla r_{i2}(\beta) \quad \ldots \quad \nabla r_{id_i}(\beta) ) \]

For each of the \( j \in [1, d_i] \) observations for the \( i^{th} \) individual, \( \nabla r_{ij}(\beta)^t \) can be formed column by column, where \( \nabla r_{ij}(\beta) \) denotes the \( j^{th} \) column. \( \nabla \mu_{ij}(\beta) \) and \( \nabla \sigma^2_{ij}(\beta) \) respectively reflect the derivative of the mean and variance of the hypothesized density, with respect to \( \beta \).

\[
\nabla r_{ij}(\beta) = -\frac{1}{\sigma_{ij}(\beta)} \nabla \mu_{ij}(\beta) - \frac{1}{2} \frac{y_{ij} - \mu_{ij}(\beta)}{\sigma^2_{ij}(\beta)} \nabla \sigma^2_{ij}(\beta) \in \mathbb{R}^p \tag{8}
\]

\[
\nabla \mu_{ij}(\beta) = \frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \frac{\partial \eta_{ij}(\beta)}{\partial \beta} = \begin{pmatrix}
\left( \frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \right) \star \frac{\partial x_{ij}(\beta)}{\partial \beta_1} \\
\vdots \\
\left( \frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \right) \star \frac{\partial x_{ij}(\beta)}{\partial \beta_p}
\end{pmatrix} = \begin{pmatrix}
\frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \\
\frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \\
\vdots \\
\frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)}
\end{pmatrix} \star \begin{pmatrix}
x_{ij1} \\
x_{ij2} \\
\vdots \\
x_{ijp}
\end{pmatrix} = \frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \star x_{ij} \in \mathbb{R}^p
\]

\[
\nabla \sigma^2_{ij}(\beta) = \frac{\partial \sigma^2_{ij}(\beta)}{\partial \mu_{ij}(\beta)} \frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \frac{\partial \eta_{ij}(\beta)}{\partial \beta} = \frac{\partial \sigma^2_{ij}(\beta)}{\partial \mu_{ij}(\beta)} \frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \star x_{ij} \in \mathbb{R}^p
\]

For the Gaussian base model, since the identity function is the appropriate canonical link, we have that \( \mu_{ij}(\beta) = \eta_{ij}(\beta) = x_{ij1} \beta_1 + \ldots + x_{ijp} \beta_p \) where \( x_{ij} \) denotes the vector of \( p \) covariate values for \( j^{th} \) measurement of the \( i^{th} \) person.

\[
\nabla \mu_{ij}(\beta) = \frac{\partial \mu_{ij}(\beta)}{\partial \beta} = \begin{pmatrix}
\frac{\partial x_{ij}(\beta)}{\partial \beta_1} \\
\vdots \\
\frac{\partial x_{ij}(\beta)}{\partial \beta_p}
\end{pmatrix} = \begin{pmatrix}
x_{ij1} \\
x_{ij2} \\
\vdots \\
x_{ijp}
\end{pmatrix} = x_{ij} \in \mathbb{R}^p
\]

\[
\nabla \sigma^2_{ij}(\beta) = \frac{\partial \sigma^2_{ij}(\beta)}{\partial \mu_{ij}(\beta)} \frac{\partial \mu_{ij}(\beta)}{\partial \eta_{ij}(\beta)} \frac{\partial \eta_{ij}(\beta)}{\partial \beta} = 0 \star x_{ij} = 0 \in \mathbb{R}^p
\]

In the table below, we derive the same quantities for the Normal, Poisson, Bernoulli and negative binomial distributions, under the appropriate canonical link function. The details of the derivation for the above table is below.

| Distribution | \( g(\mu_{ij}(\beta)) = \eta_{ij}(\beta) \) | \( \mu_{ij}(\beta) \in \mathbb{R} \) | \( \sigma^2_{ij}(\beta) \in \mathbb{R}^p \) | \( \nabla \mu_{ij}(\beta) \in \mathbb{R}^p \) | \( \nabla \sigma^2_{ij}(\beta) \in \mathbb{R}^p \) |
|--------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| Normal       | Identity Link                   | \( \eta_{ij}(\beta) \)         | \( \sigma_{ij}(\beta) \)      | \( \partial \eta_{ij}(\beta) \) | \( \partial \sigma_{ij}(\beta) \) |
| Poisson      | Log Link                        | \( e^{\eta_{ij}(\beta)} \)    | \( \mu_{ij}(\beta) \)         | \( e^{\eta_{ij}(\beta)} \star x_{ij} \) | \( e^{\eta_{ij}(\beta)} \star x_{ij} \) |
| Bernoulli    | Logit Link                      | \( \frac{e^{\eta_{ij}(\beta)}}{1 + e^{\eta_{ij}(\beta)}} \) | \( \frac{e^{\eta_{ij}(\beta)}}{1 + e^{\eta_{ij}(\beta)}} \star x_{ij} \) | \( \frac{e^{\eta_{ij}(\beta)}}{1 + e^{\eta_{ij}(\beta)}} \star x_{ij} \) | \( \frac{e^{\eta_{ij}(\beta)}}{1 + e^{\eta_{ij}(\beta)}} \star x_{ij} \) |
| Negative Binomial | Log Link              | \( e^{\eta_{ij}(\beta)} \)    | \( e^{\eta_{ij}(\beta)} \star (1 + e^{\eta_{ij}(\beta)}) \) | \( e^{\eta_{ij}(\beta)} \star x_{ij} \) | \( e^{\eta_{ij}(\beta)} \star x_{ij} \) |

45
11.3.2 MM-Algorithm to Update Variance Components

To update the variance components \( \theta = \{ \theta_k, k \in [1, m] \} \), the relevant part of the loglikelihood is

\[
f(\theta) = \sum_{i=1}^{n} \ln(1 + \theta^t b_i) - \sum_{i=1}^{n} \ln(1 + \theta^t c_i) \tag{9}
\]

by defining the vectors \( b_i \) and \( c_i \) with nonnegative components

\[
b_{ik} = \frac{1}{2} r_i(\beta)^t \Omega_{ik} r_i(\beta)
\]

\[
c_{ik} = \frac{1}{2} \text{tr}(\Omega_{ik}).
\]

If \( f \) is a convex function of a vector \( x \), \( f(x) \geq f(x_0) + (x - x_0) \nabla f(x_0) \)

\[
\sum_{i=1}^{n} -\ln(1 + \theta^t c_i) \geq -\sum_{i=1}^{n} \ln(1 + \theta^t c_i) + (1 + \theta^t c_i) * \frac{-1}{1 + \theta^t c_i}
\]

\[
= c_{1}^{(t)} - \sum_{i=1}^{n} \frac{1}{1 + \theta^t c_i}(\theta^t c_i - \theta^t c_i)
\]

\[
= -\sum_{i=1}^{n} \frac{1}{1 + \theta^t c_i}(\theta^t c_i) + c^{(t)}
\]

The first term is minorized by the Jensen’s function

\[
\sum_{i=1}^{n} \left[ \frac{1}{1 + \theta^t b_i} \ln \left( \frac{1 + \theta^t b_i}{1} \right) + \sum_{k=1}^{m} \frac{\theta_{rk} b_{ik}}{1 + \theta^t b_i} \ln \left( \frac{1 + \theta^t b_i}{\theta_{rk} b_{ik}} \right) \right].
\]

Proof: \( f(a) \) is a concave vector function if for any vectors \( a_1, a_2, \lambda \in [0, 1] \)

\[
f(\lambda a_1 + (1 - \lambda) a_2)
\]

\[
= \frac{1}{1 + \theta^t b_i}(1 + \theta^t b_i) + \frac{\theta^t b_i}{1 + \theta^t b_i} \frac{1 + \theta^t b_i}{\theta^t b_i}
\]

\[
= \frac{1 + \theta^t b_i}{1 + \theta^t b_i} \ln(1 + \theta^t b_i) + \frac{\theta^t b_i}{1 + \theta^t b_i} \ln(1 + \theta^t b_i)
\]

46
\[
\sum_{i=1}^{n} \ln(1 + \theta_i b_i) \geq \sum_{i=1}^{n} \left[ \frac{1}{1 + \theta_i b_i} \ln \left( \frac{1 + \theta_i b_i}{1} \right) + \sum_{k=1}^{m} \frac{\theta_{rk} b_{ik}}{1 + \theta_i b_i} \ln \left( \frac{1 + \theta_i b_i}{\theta_{rk} b_{ik} \theta_{r+1,k} b_{ik}} \right) \right].
\]

The sum of these two minorizations constitutes the surrogate \( h(a \mid a_r) \).

\[
h(\theta \mid \theta_r) = -\sum_{i=1}^{n} \frac{1}{1 + \theta_i c_i} (\theta_i c_i) + \left[ \frac{1}{1 + \theta_i b_i} \ln \left( \frac{1 + \theta_i b_i}{1} \right) + \sum_{k=1}^{m} \theta_{rk} b_{ik} \ln \left( \frac{1 + \theta_i b_i}{\theta_{rk} b_{ik} \theta_{r+1,k} b_{ik}} \right) \right].
\]

We can maximize the surrogate function by taking a derivative with respect to \( a_k, k \in [1, m] \). The stationarity condition \( \nabla h(a \mid a_r) = 0 \) has components

\[
\frac{\partial}{\partial \theta_k} h(\theta \mid \theta_r) = \sum_{i=1}^{n} \frac{b_{ik}}{1 + \theta_i b_i} \frac{1}{1 + \theta_i b_i} \theta_{rk} b_{ik} \theta_{r+1,k} b_{ik} = 0
\]

with solution

\[
\theta_{r+1,k} = \theta_{rk} \left( \frac{\sum_{i=1}^{n} b_{ik}}{\sum_{i=1}^{n} \frac{b_{ik}}{1 + \theta_i b_i}} \right).
\]

### 11.3.3 Quasi-Newton Algorithm

Alternatively, we can estimate the mean and variance parameters jointly using the Quasi-Newton algorithm.

**Score and Hessian** For the AR(1) model, \( \theta = \{\sigma^2, \rho\} \), the score (gradient of loglikelihood function) is

\[
\nabla_{\sigma^2} L = -\sum_{i=1}^{n} \frac{d_i}{1 + \frac{d_i}{2} \sigma^2} + \sum_{i=1}^{n} \frac{1}{2} r_i(\beta)^t V_i(\rho) r_i(\beta)
\]

\[
\nabla_{\rho} L = \sum_{i=1}^{n} \frac{1}{1 + \frac{\sigma^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta)} \frac{\sigma^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta).
\]

The approximate Hessian is

\[
d^2_{\sigma^2} L = \sum_{i=1}^{n} \frac{(\frac{d_i}{2})^2}{(1 + \frac{d_i}{2} \sigma^2)^2} - \sum_{i=1}^{n} \frac{(\frac{1}{2} r_i(\beta)^t V_i(\rho) r_i(\beta))^2}{(1 + \frac{\sigma^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta))^2}
\]

\[
d^2_{\rho} L = \sum_{i=1}^{n} \frac{1}{1 + \frac{\sigma^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta)} \frac{\sigma^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta)
\]

\[
- \sum_{i=1}^{n} \frac{1}{(1 + \frac{\sigma^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta))^2} \left( \frac{\sigma^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta) \right)^2
\]

47
For the CS model, $\theta = (\sigma^2, \rho)$, and the gradient is
\[
\nabla_{\sigma^2} L = -\sum_{i=1}^{n} \frac{d_i^2}{1 + d_i \sigma^2} + \sum_{i=1}^{n} \frac{1}{2} \frac{r_i(\beta)^t V_i(\rho) r_i(\beta)}{1 + \frac{\alpha^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta)}
\]
\[
\nabla_{\rho} L = \sum_{i=1}^{n} \frac{1}{1 + \frac{\alpha^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta)} * \frac{\sigma^2}{2} r_i(\beta)^t \nabla V_i(\rho) r_i(\beta).
\]

The approximate Hessian is
\[
d_{\sigma^2}^2 L = \sum_{i=1}^{n} \frac{(d_i^2)^2}{(1 + \frac{d_i}{2} \sigma^2)^2} - \sum_{i=1}^{n} \frac{(\frac{1}{2} r_i(\beta)^t V_i(\rho) r_i(\beta))^2}{(1 + \frac{\alpha^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta))^2}
\]
\[
d_{\rho}^2 L = -\sum_{i=1}^{n} \frac{1}{(1 + \frac{\alpha^2}{2} r_i(\beta)^t V_i(\rho) r_i(\beta))^2} * \left( \frac{\sigma^2}{2} r_i(\beta)^t \nabla V_i(\rho) r_i(\beta) \right)^2,
\]

where $\nabla V_i(\rho)$ and $\nabla^2 V_i(\rho)$ are, respectively, the element-wise first and second derivatives of the matrix $V_i(\rho)$ with respect to $\rho$.

For the VM model the gradient is
\[
\nabla_{\theta} f(\theta) = \sum_{i=1}^{n} \frac{1}{1 + \theta^t b_i} * b_i - \sum_{i=1}^{n} \frac{1}{1 + \theta^t c_i} * c_i.
\]

The approximate Hessian is
\[
\nabla_{\theta, \theta}^2 f(\theta) = -\sum_{i=1}^{n} \frac{1}{(1 + \theta^t b_i)^2} * b_i b_i^t + \sum_{i=1}^{n} \frac{1}{(1 + \theta^t c_i)^2} * c_i c_i^t.
\]

### 11.3.4 Negative Binomial

**Estimating Nuisance Parameter** To estimate the nuisance parameter $r$ in a Negative Binomial model, we use maximum likelihood. Because we are dealing with 1 parameter optimization, Newton’s method is a good candidate due to its quadratic rate of convergence. The full loglikelihood is
\[
-\sum_{i=1}^{n} \ln \left( 1 + \frac{1}{2} \text{tr}(\Gamma_i) \right) + \sum_{i=1}^{n} \sum_{j=1}^{d_i} \ln f_{ij}(y_{ij} \mid \beta) + \sum_{i=1}^{n} \ln \left( 1 + \frac{1}{2} r_i(\beta)^t \Gamma_i r_i(\beta) \right)
\]
where only the 2nd and 3rd term depends on \(r\). First consider the 2nd term. Because

\[ \mu_{ij} = \frac{r(1-p_{ij})}{p_{ij}}, \quad p_{ij} = \frac{r}{r+\mu_{ij}}, \]

the 2nd term of the loglikelihood is

\[
\sum_{i=1}^{n} \sum_{j=1}^{d_{i}} \ln \left[ \left( \frac{y_{ij} + r - 1}{y_{ij}} \right) p_{ij}^r (1-p_{ij})^{y_{ij}} \right]
= \sum_{i=1}^{n} \sum_{j=1}^{d_{i}} \ln \left( y_{ij} + r - 1 \right) + r \ln \left( \frac{r}{\mu_{ij} + r} \right) + y_{ij} \ln \left( \frac{\mu_{ij}}{\mu_{ij} + r} \right)
= \sum_{i=1}^{n} \sum_{j=1}^{d_{i}} \ln((y_{ij} + r - 1)!) - \ln(y_{ij}!) - \ln((r-1)!) + r \ln(r) - (r + y_{ij}) \ln(\mu_{ij} + r) + y_{ij} \ln(\mu_{ij})
\]

Let \(\Psi^{(0)}\) be the digamma function and \(\Psi^{(1)}\) the trigamma function, then the first and second derivative is

\[
\sum_{i=1}^{n} \sum_{j=1}^{d_{i}} \Psi^{(0)} (y_{ij} + r) - \Psi^{(0)} (r) + 1 + \ln(r) - \frac{r + y_{ij}}{\mu_{ij} + r} - \ln(\mu_{ij} + r),
\]

\[
\sum_{i=1}^{n} \sum_{j=1}^{d_{i}} \Psi^{(1)} (y_{ij} + r) - \Psi^{(0)} (r) + \frac{1}{r} - \frac{2}{\mu_{ij} + r} + \frac{r + y_{ij}}{(\mu_{ij} + r)^2}.
\]

Now consider the 3rd term of the full loglikelihood. First recall

\[ D_i = \text{diagonal}(\sqrt{\text{var}(y_i)}) \]
\[ \text{var}(y_{ij}) = \frac{r(1-p_{ij})}{p_{ij}^2} = \frac{e^{h_{ij}}(e^{h_{ij}} + r)}{r}. \]

Using multiple chain rules,

\[
\frac{d}{dr} \ln \left( 1 + \frac{1}{2} r_i(\beta)^T \Gamma_i r_i(\beta) \right) = \sum_{i=1}^{n} \frac{r_i(\beta)^T \Gamma_i d r_i}{1 + \frac{1}{2} r_i(\beta)^T \Gamma_i r_i(\beta)}
\]
\[
\frac{d^2}{dr^2} \ln \left( 1 + \frac{1}{2} r_i(\beta)^T \Gamma_i r_i(\beta) \right) = \sum_{i=1}^{n} \frac{-[r_i(\beta)^T \Gamma_i d r_i]^2}{[1 + \frac{1}{2} r_i(\beta)^T \Gamma_i r_i(\beta)]^2} + \frac{dr(\beta)^T \Gamma_i d r_i(\beta) + r(\beta)^T \Gamma_i d r_i^2(\beta)}{1 + \frac{1}{2} r_i(\beta)^T \Gamma_i r_i(\beta)}
\]

49
where
\[ r_i(\beta) = D_i^{-1}(y_i - \mu_i) \]
\[ dr_i(\beta) = -D_i^{-1}dD_iD_i^{-1}(y_i - \mu_i) \]
\[ dr_i^2(\beta) = [2D_i^{-1}dD_iD_i^{-1}dD_iD_i^{-1} - D_i^{-1}d^2D_iD_i^{-1}](y_i - \mu_i) \]
\[ dD_i = \text{diagonal}\left(\frac{d}{dr}\sqrt{\frac{e^{\eta_j}(e^{\eta_j} + r)}{r}}\right) = \text{diagonal}\left(\frac{-e^{2\eta_j}}{2r^{1.5}\sqrt{e^{\eta_j}(e^{\eta_j} + r)}}\right) \]
\[ d^2D_i = \text{diagonal}\left(\frac{e^{3\eta}}{4r^{1.5}(e^{\eta_j} + r)^{1.5}} + \frac{3e^{2\eta}}{4r^{2.5}(e^{\eta_j} + r)^{0.5}}\right). \]

Note we used the identity \( df(X)^{-1} = -f(X)^{-1}df(X)f(X)^{-1} \) for obtaining \( dr_i(\beta) \) and for obtaining \( dr_i^2(\beta) \), chain rule implies
\[ d(f(X)^{-1}) = [f(X)^{-1}df(X)f(X)^{-1}]df(X)f(X)^{-1} + f(X)^{-1}d(df(X)f(X)^{-1}) \]
\[ = -f(X)^{-1}df(X)f(X)^{-1}df(X)f(X)^{-1} + f(X)^{-1}[df(X)(-f(X)^{-1}df(X)f(X)^{-1}) + d^2f(X)f(X)^{-1}] \]
\[ = -2f(X)^{-1}df(X)f(X)^{-1}df(X)f(X)^{-1} + f(X)^{-1}d^2f(X)f(X)^{-1} \]

In summary, we update the nuisance parameter \( r \) using Newton's update
\[ r_{n+1} = r_n - \frac{\frac{d}{dr}L(r \mid \mu, \Gamma, y)}{\frac{d^2}{dr^2}L(r \mid \mu, \Gamma, y)} \]
where
\[ \frac{d}{dr}L(r \mid \mu, \Gamma, y) = \sum_{i=1}^{n} \sum_{j=1}^{d_i} \Psi^{(0)}(y_{ij} + r) - \Psi^{(0)}(r) + 1 + \ln(r) - \frac{r + y_{ij}}{\mu_{ij} + r} - \ln(\mu_{ij} + r) \]
\[ + \sum_{i=1}^{n} \frac{r_i(\beta)^T \Gamma_i dr_i}{1 + \frac{1}{2}r_i(\beta)^T \Gamma_i r_i(\beta)} \]
\[ \frac{d^2}{dr^2}L(r \mid \mu, \Gamma, y) = \sum_{i=1}^{n} \sum_{j=1}^{d_i} \Psi^{(1)}(y_{ij} + r) - \Psi^{(0)}(r) + \frac{1}{r} - \frac{2}{\mu_{ij} + r} + \frac{r + y_{ij}}{(\mu_{ij} + r)^2} \]
\[ - \sum_{i=1}^{n} \frac{[r_i(\beta)^T \Gamma_i dr_i]^2}{[1 + \frac{1}{2}r_i(\beta)^T \Gamma_i r_i(\beta)]^2} + \frac{dr_i(\beta)^T \Gamma_i dr_i(\beta) + r_i(\beta)^T \Gamma_i dr_i^2(\beta)}{1 + \frac{1}{2}r_i(\beta)^T \Gamma_i r_i(\beta)} \]

50
For stability, we need to (1) perform line-search and (2) set the second derivative equal to 1 if it is negative. By default, we allow for a maximum of 10 block iterations; in each block iteration, we allow for a maximum of 15 iterations for the quasi-newton update of $\beta$ and $\theta$, and a maximum of 10 newton iterations for the update of $r$.

### 11.3.5 Compound Symmetric Covariance:

Under the Compound Symmetric (CS) parameterization of $\Gamma_i$,

$$
\Gamma_i = \sigma^2 \times \left[ \rho \mathbf{1}_{d_i} \mathbf{1}_{d_i}^T + (1 - \rho) \mathbf{I}_{d_i} \right] = \sigma^2 \times \mathbf{V}_i(\rho)
$$

**Bounding Correlation Parameter**  To ensure that the covariance matrix $\Gamma_i$ is positive semi-definite, we will focus on $\mathbf{V}_i(\rho)$ and use an eigenvalue argument to bound $\rho \in (-\frac{1}{d_i-1}, 1)$. Let $\mathbf{v}$ be a vector of dimension $d_i$ such that $\langle \mathbf{v}, \mathbf{v} \rangle = 1$. We will find the conditions on $\rho$ such that $\mathbf{v}^T \mathbf{V}_i(\rho) \mathbf{v} \geq 0$.

$$
\mathbf{v}^T \mathbf{V}_i(\rho) \mathbf{v} = \mathbf{v}^T \left[ \rho \mathbf{1}_{d_i} \mathbf{1}_{d_i}^T + (1 - \rho) \mathbf{I}_{d_i} \right] \mathbf{v}
= \rho \mathbf{v}^T \mathbf{1}_{d_i} \mathbf{1}_{d_i}^T \mathbf{v} + (1 - \rho) \mathbf{v}^T \mathbf{v}
= \rho (\mathbf{1}_{d_i}^T \mathbf{v})^2 + 1 - \rho
= \rho \left( (\mathbf{1}_{d_i}^T \mathbf{v})^2 - 1 \right) + 1
\geq 0
$$

Now solving for $\rho$ and using the Cauchy-Schwartz Inequality, we get

$$
\rho \geq \frac{-1}{(\mathbf{1}_{d_i}^T \mathbf{v})^2 - 1}
\geq \frac{-1}{\mathbf{1}_{d_i}^T \mathbf{1}_{d_i} \ast (\mathbf{v}^T \mathbf{v}) - 1}
= \frac{-1}{d_i - 1}
$$
11.4 Gaussian Base

This section considers the special case of Gaussian base in the quasi-copula framework, and presents detailed derivations of data generation and estimation methods. The joint density of \( y \in \mathbb{R}^d \) is

\[
\left( c + \frac{1}{2} \text{tr}(\Gamma) \right)^{-1} \left( \frac{1}{\sqrt{2\pi \sigma_0^2}} \right)^d e^{-\frac{\|y-\mu\|^2}{2\sigma_0^2}} \left[ c + \frac{1}{2\sigma_0^2}(y - \mu)^T \Gamma (y - \mu) \right].
\]

The parameter \( c \geq 0 \) tips the balance between the independent and dependent components.

11.4.1 Moments

In the Gaussian case, we have

\[
\begin{align*}
\mathbb{E}(y_i) &= \mu_i \\
\text{Var}(y_i) &= \sigma_0^2 \left( 1 + \frac{\gamma_{ii}}{c + \frac{1}{2} \text{tr}(\Gamma)} \right) \\
\text{Cov}(y_i, y_j) &= \sigma_0^2 \frac{\gamma_{ij}}{c + \frac{1}{2} \text{tr}(\Gamma)}, \\
\text{Cor}(y_i, y_j) &= \frac{\gamma_{ij}}{\sqrt{(c + \frac{1}{2} \text{tr}(\Gamma)) + \gamma_{ii})(c + \frac{1}{2} \text{tr}(\Gamma) + \gamma_{jj})}}.
\end{align*}
\]

In summary,

\[
\text{Cov}(y) = \sigma_0^2 \left[ \mathbf{I} + \left( \frac{1}{c + \frac{1}{2} \text{tr}(\Gamma)} \right) \Gamma \right].
\]

In the special case of \( \Gamma = \sigma_1^2 \mathbf{I} \), we have

\[
\begin{align*}
\text{Var}(y_i) &= \sigma_0^2 \left( 1 + \frac{\sigma_1^2}{c + \frac{\sigma_1^2}{2\sigma_0^2}} \right) \mathbf{I} \\
\text{Cov}(y_i, y_j) &= 0, \quad i \neq j.
\end{align*}
\]

In the regression model, we would keep the variance \( \sigma_0^2 \) parameter for more flexibility in modeling the variance.

Random number generation  If we are able to generate a residual vector \( R \) from the (standardized) Gaussian copula model

\[
\left[ 1 + \frac{1}{2} \text{tr}(\Gamma) \right]^{-1} \left( \frac{1}{\sqrt{2\pi}} \right)^d e^{-\frac{1}{2}r^T r} \left( 1 + \frac{1}{2} r^T \Gamma r \right),
\]

\[
52
\]
then $Y = \sigma_0 R + \mu$ is a desired sample from density $f^{10}$.

To generate a sample from the standardized Gaussian copula model, we first sample sequentially from the conditional distributions $R_k \mid R_1, \ldots, R_{k-1}$ for $k = 2, \ldots, d$.

- To generate $R_1$ from its marginal density

\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{r_1^2}{2}} \left( 1 + \frac{\gamma_{11} r_1^2}{2} + \frac{1}{2} \sum_{i=2}^d \gamma_{ii} \right),
\]

we recognize it as a mixture of three distributions $\text{Normal}(0, 1)$, $\sqrt{\frac{\gamma_{11}}{2}}$ and $-\sqrt{\frac{\gamma_{11}}{2}}$ with mixing probabilities $\frac{1 + 0.5 \sum_{i=2}^d \gamma_{ii}}{1 + 0.5 \sum_{i=1}^d \gamma_{ii}}$, $\frac{0.25 \gamma_{11}}{1 + 0.5 \sum_{i=1}^d \gamma_{ii}}$ and $\frac{0.25 \gamma_{11}}{1 + 0.5 \sum_{i=1}^d \gamma_{ii}}$ respectively.

- Next we consider generating $R_2$ from the conditional distribution $R_2 \mid R_1$. Dividing the marginal distribution of $(R_1, R_2)$

\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{r_1^2}{2}} \left( 1 + \frac{\gamma_{12} r_1 r_2}{2} + \frac{\gamma_{22} r_2^2}{2} + \frac{\gamma_{11} r_1^2}{2} + \frac{1}{2} \sum_{i=3}^d \gamma_{ii} \right)
\]

by the marginal distribution of $R_1$, (11) yields the conditional density

\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{r_1^2}{2}} \left( 1 + \frac{\gamma_{11} r_1^2}{2} + \frac{1}{2} \sum_{i=2}^d \gamma_{ii} + \frac{\gamma_{12} r_1 r_2}{2} + \frac{\gamma_{22} r_2^2}{2} \right)
\]

which unfortunately is not a mixture of standard distributions. However we can evaluate its cumulative distribution function (CDF)

\[
F(x) = \Phi(x) - \gamma_{12} r_1 \phi(x) + \frac{\gamma_{22}}{2} \left[ 1 + \frac{\text{sgn}(x)}{2} F_{\chi_3^2}(x^2) \right] \left( 1 + \frac{\gamma_{11} r_1^2}{2} + \frac{1}{2} \sum_{i=2}^d \gamma_{ii} \right)
\]

in terms of the density $\phi$ and CDF $\Phi$ of standard normal and the CDF $F_{\chi_3^2}$ of chi-squared distribution with degree of freedom 3. This suggests the inverse CDF approach. To generate one sample from $R_2 \mid R_1$, we draw a uniform variate $U$ and use nonlinear root finding to locate $R_2$ such that $F(R_2) = U$.

- In general, the conditional distribution $R_k \mid R_1, \ldots, R_{k-1}$ has density

\[
\frac{1}{\sqrt{2\pi}} e^{-\frac{r_k^2}{2}} \left( 1 + \frac{1}{2} r_{[k-1]}^T \Gamma_{[k-1], [k-1]} r_{[k-1]} + \frac{1}{2} \sum_{i=k+1}^d \gamma_{ii} + (\sum_{i=1}^{k-1} r_i \gamma_{ik}) r_k + \frac{\gamma_{kk} r_k^2}{2} \right)
\]

\[
1 + \frac{1}{2} r_{[k-1]}^T \Gamma_{[k-1], [k-1]} r_{[k-1]} + \frac{1}{2} \sum_{i=k}^d \gamma_{ii}
\]

53
and CDF

\[
\left(1 + \frac{1}{2}r_T^{T}[k-1] \Gamma_{[k-1],[k-1]} r_{[k-1]} + \frac{1}{2} \sum_{i=k+1}^{d} r_{i}\gamma_{ii} \right) \Phi(x) - (\sum_{i=1}^{k-1} r_i\gamma_{ik}) \phi(x) + \frac{\gamma_{kk}}{2} \left[ \frac{1}{2} + \frac{\text{sgn}(x)}{2} F_{\chi^2_2}(x^2) \right].
\]

We apply the inverse CDF approach to sample \( R_k \) given \( R_1, \ldots, R_{k-1} \).

For a general GLM model, we need to sample from conditional densities of form \( cf(y)(a_0 + a_1 y + a_2 y^2) \) where \( a_i, \ i = 1, 2, 3 \), are constants and \( c \) is a normalizing constant. For most continuous distributions, e.g., exponential, gamma, beta, chi-squared, and beta, the CDF can be expressed conveniently using special functions.

### 11.4.2 Parameter Estimation

Suppose we have \( n \) independent realizations \( y_i \) from the quasi-copula density. Each of these may be of different dimensions, \( d_i \). Assuming the component distribution \( y_i \sim \text{Normal}(X_i\beta, \sigma_0^2 I_{d_i}) \), the component densities take form

\[
\ln f_i(y_i | \beta, \sigma_0^2) = -\frac{d_i}{2} \ln 2\pi - \frac{d_i}{2} \ln \sigma_0^2 - \frac{1}{2} \frac{\|y_i - X_i\beta\|^2}{\sigma_0^2}
\]

and the joint loglikelihood of the sample is

\[
-\sum_i \ln \left( c + \frac{1}{2} \text{tr}(\Gamma_i) \right) - \sum_i \frac{d_i}{2} \ln 2\pi - \sum_i \frac{d_i}{2} \ln \sigma_0^2 - \frac{1}{2} \sum_i \frac{\|y_i - X_i\beta\|^2}{\sigma_0^2}
+ \sum_i \ln \left[ c + \frac{1}{2\sigma_0^2} (y_i - X_i\beta)^T \Gamma_i (y_i - X_i\beta) \right]
= -\sum_i \ln \left( c + \frac{1}{2} \text{tr}(\Gamma_i) \right) - \sum_i \frac{d_i}{2} \ln 2\pi + \sum_i \frac{d_i}{2} \ln \tau - \frac{\tau}{2} \sum_i \frac{\|y_i - X_i\beta\|^2}{\sigma_0^2}
+ \sum_i \ln \left[ c + \frac{\tau}{2} (y_i - X_i\beta)^T \Gamma_i (y_i - X_i\beta) \right]
\]

where \( \Gamma_i = \sum_{k=1}^{m} \theta_k V_{ik} \) are parameterized via variance components \( \theta = (\theta_1, \ldots, \theta_m) \). We work with the parameterization \( \tau = \sigma_0^{-2} \) because the loglikelihood is concave in \( \tau \).
Score and Hessian  The score (gradient of loglikelihood function) is

\[
\nabla \beta = \sigma_0^{-2} \sum_i X_i^T (y_i - X_i \beta) - \sum_i \frac{X_i^T \Gamma_i (y_i - X_i \beta)}{c \sigma_0^2 + \frac{1}{2} (y_i - X_i \beta)^T \Gamma_i (y_i - X_i \beta)} \\
= \tau \sum_i X_i^T (y_i - X_i \beta) - \tau \sum_i \frac{X_i^T \Gamma_i (y_i - X_i \beta)}{c + \frac{\tau}{2} (y_i - X_i \beta)^T \Gamma_i (y_i - X_i \beta)} \\
\nabla \tau = \frac{\sum_i d_i}{2\tau} - \frac{1}{2} \sum_i \|y_i - X_i \beta\|_2^2 + \sum_i \frac{1}{c + \frac{\tau}{2} (y_i - X_i \beta)^T \Gamma_i (y_i - X_i \beta)} \\
\nabla c = \sum_i \frac{1}{c + \frac{\tau}{2} (y_i - X_i \beta)^T \Gamma_i (y_i - X_i \beta)} \\
\nabla \theta = -\sum_i \left( c + \sum_k \theta_k t_{ik} \right)^{-1} t_i + \tau \sum_i \left( c + \sum_k \theta_k q_{ik} \right)^{-1} q_i
\]

where

\[
\begin{align*}
    t_{ik} & = \frac{1}{2} \text{tr} (\Omega_{ik}) , \quad t_i = (t_{i1}, \ldots, t_{im})^T \\
    q_{ik} & = \frac{1}{2} (y_i - X_i \beta)^T \Omega_{ik} (y_i - X_i \beta) , \quad q_i = (q_{i1}, \ldots, q_{im})^T 
\end{align*}
\]
The Hessian is

\[
\nabla^2_{\beta,\beta} = -\tau \sum_i X_i^T X_i + \sum_i \frac{\tau X_i^T \Gamma_i X_i}{c + \frac{\tau}{2} (y_i - \mathbf{x}_i \beta)^T \Gamma_i (y_i - \mathbf{x}_i \beta)} \\
- \sum_i \frac{\tau^2 [X_i^T \Gamma_i (y_i - \mathbf{x}_i \beta)][X_i^T \Gamma_i (y_i - \mathbf{x}_i \beta)]^T}{[c + \frac{\tau}{2} (y_i - \mathbf{x}_i \beta)^T \Gamma_i (y_i - \mathbf{x}_i \beta)]^2} \\
\approx -\tau \sum_i X_i^T X_i - \tau \sum_i \frac{\tau [X_i^T \Gamma_i (y_i - \mathbf{x}_i \beta)][X_i^T \Gamma_i (y_i - \mathbf{x}_i \beta)]^T}{[c + \frac{\tau}{2} (y_i - \mathbf{x}_i \beta)^T \Gamma_i (y_i - \mathbf{x}_i \beta)]^2}
\]

\[
\nabla^2_{\beta,\tau} = \sum_i X_i^T (y_i - \mathbf{x}_i \beta) - \sum_i \frac{X_i^T \Gamma_i (y_i - \mathbf{x}_i \beta)}{[c + \frac{\tau}{2} (y_i - \mathbf{x}_i \beta)^T \Gamma_i (y_i - \mathbf{x}_i \beta)]^2}
\]

\[
\nabla^2_{\beta,\theta} = \sum_i \frac{X_i^T \Gamma_i (y_i - \mathbf{x}_i \beta) q_i^T}{[c + \frac{\tau}{2} (y_i - \mathbf{x}_i \beta)^T \Gamma_i (y_i - \mathbf{x}_i \beta)]^2}
\]

\[
\nabla^2_{\tau,\tau} = -\frac{\sum_i d_i}{2 \tau^2} - \sum_i \left[ \frac{1}{2} (y_i - \mathbf{x}_i \beta)^T \Gamma_i (y_i - \mathbf{x}_i \beta) \right]^2 \\
\nabla^2_{\tau,\theta} = -\sum_i \frac{X_i^T \Gamma_i (y_i - \mathbf{x}_i \beta) q_i^T}{[c + \frac{\tau}{2} (y_i - \mathbf{x}_i \beta)^T \Gamma_i (y_i - \mathbf{x}_i \beta)]^2}
\]

\[
\nabla^2_{\theta,\theta} = \sum_i \left( c + \sum_k \theta_k t_{ik} \right)^{-2} t_i^T t_i - \tau \sum_i \left( c + \sum_k \theta_k q_{ik} \right)^{-2} q_i q_i^T.
\]

Note \(E[\nabla^2_{\beta,\tau}], E[\nabla^2_{\beta,\theta}],\) and \(E[\nabla^2_{\tau,\theta}]\) are approximately zero.

**MM algorithm** Because the MM update of \(\theta\) and \(\tau\) is cheap, we maximize the profiled likelihood. That is, after each Newton update of \(\beta\), we update \((\tau, \theta)\) conditional on current \(\beta\) using the MM algorithm and evaluate the gradient and (approximate) Hessian using the newest \((\tau, \theta)\). To update \(\tau\) and \(\theta\) given \(\beta\), the relevant objective function is

\[
- \sum_i \ln \left( c + \sum_k \theta_k t_{ik} \right) + \frac{\sum_i d_i}{2} \ln \tau - \frac{\sum_i y_i^2}{2} \tau + \sum_i \ln \left( c + \tau \sum_k \theta_k q_{ik} \right),
\]

56
which is minorized by

\[- \sum_i \sum_k \frac{t_{ik}}{c(t)} + \sum_k \theta_k^{(t)} t_{ik} - \sum_i \frac{1}{c(t)} + \sum_k \theta_k^{(t)} t_{ik} \cdot c\]

\[+ \frac{\sum_i d_i}{2} \ln \tau - \frac{\sum_i r_i^2}{2} \cdot \tau\]

\[+ \sum_i \sum_k \frac{\tau(t) \theta_k^{(t)} q_{ik}}{c(t) + \tau(t) \cdot \sum_k \theta_k^{(t)} q_{ik}} \cdot (\ln \tau + \ln \theta_k)\]

\[+ \sum_i \frac{c(t) + \tau(t) \cdot \sum_k \theta_k^{(t)} q_{ik}}{c(t)} \ln c\]

\[+ \text{const.}\]

The resultant updates are

\[\tau^{(t+1)} = \frac{\sum_i d_i + 2 \sum_i \frac{\tau(t) q_i^{(t)}}{c(t) + \tau(t) q_i^{(t)}}}{\sum_i r_i^2}\]

\[c^{(t+1)} = c(t) \cdot \frac{\sum_i \frac{1}{c(t) + \tau(t) q_i^{(t)}}}{\sum_i \frac{1}{c(t) + \tau(t) q_i^{(t)}}}\]

\[\theta_k^{(t+1)} = \theta_k^{(t)} \cdot \frac{\sum_i \frac{\tau(t) q_{ik}}{t_{ik}}}{\sum_i \frac{t_{ik}}{c(t) + \tau(t) q_i^{(t)}}}, \quad k = 1, \ldots, m,\]

where \(q_i^{(t)} = \sum_k \theta_k^{(t)} q_{ik}\) and \(t_i^{(t)} = \sum_k \theta_k^{(t)} t_{ik}\).

If we opt to use the optimal quadratic minorization

\[\ln(1 + x) \geq \ln(1 + x^{(t)}) + (x - x^{(t)}) - \frac{x^2 - x^{2(t)}}{2(1 + x^{(t)})},\]

the minorization function becomes

\[- \sum_i \sum_k \frac{t_{ik}}{1 + \sum_k \theta_k^{(t)} t_{ik}} \cdot \theta_k + \frac{\sum_i d_i}{2} \ln \tau + \left(\sum_i \sum_k \theta_k q_{ik} - \frac{\sum_i r_i^2}{2}\right) \cdot \tau - \frac{\tau^2}{2} \cdot \sum_i \frac{(\sum_k \theta_k q_{ik})^2}{1 + \tau(t) \cdot \sum_k \theta_k^{(t)} q_{ik}}\]
To update $\tau$ given $\theta_k$, let

$$\begin{align*}
a(t) &= \sum_i \frac{(\sum_k \theta_k(t) q_{ik})^2}{1 + \tau(t) \sum_k \theta_k(t) q_{ik}}, \\
b(t) &= \sum_i \sum_k \theta_k(t) q_{ik} - \frac{\sum_i r_i^2}{2} \\
c(t) &= \frac{\sum_i d_i}{2}
\end{align*}$$

then

$$\tau(t+1) = \frac{b(t) + \sqrt{b^2(t) + 4a(t)c(t)}}{2a(t)}.$$ 

To update $\theta_k$ given $\tau$, we minimize quadratic function

$$\frac{1}{2} \sigma^2 T \mathbf{Q}^T \mathbf{W}(t) \mathbf{Q} \sigma^2 - \mathbf{c}^T(t) \sigma^2$$

subject to nonnegativity constraint $\theta_k \geq 0$, where $\mathbf{W}(t) = \text{diag}(w_1(t), \ldots, w_n(t))$ with

$$w_i(t) = \frac{\tau^2(t)}{1 + \tau(t) \sum_k \theta_k(t) q_{ik}}$$

and $\mathbf{c}(t)$ has entries

$$c_{k}^{(t)} = \tau(t) \sum_i q_{ik} - \sum_i \frac{t_{ik}}{1 + \sum_k \theta_k(t) t_{ik}}.$$ 

It turns out this update based on quadratic minorization converges slower than the update based on Jensen’s inequality.

### 11.5 Additional Simulation Study Results

In each simulation scenario, the non-intercept entries of the predictor matrix $X_i$ are independent standard normal deviates. When simulating under our model for the CS and AR(1) covariance structures, the true regression coefficients $\beta_{\text{true}} \sim \text{Uniform}(-2, 2)$. When comparing estimates with MixedModels.jl under the random intercept model for the Poisson, Bernoulli and negative binomial base, smaller regression coefficients $\beta_{\text{true}} \sim \text{Uniform}(-0.2, 0.2)$ hold. For Gaussian base, all precisions $\tau_{\text{true}} = 100$. For the negative binomial base, all dispersion parameters are
$r_{\text{true}} = 10$. Under both CS and AR(1) parameterizations of $\Gamma_i$, $\sigma^2_{\text{true}} = 0.5$ and $\rho_{\text{true}} = 0.5$. Each simulation scenario was run on 100 replicates for each sample size $n \in \{100, 1000, 10000\}$ and number of observations $d_i \in \{2, 5, 10, 15, 20, 25\}$ per independent sampling unit. By default, convergence tolerances are set to $10^{-6}$.

Under the VC parameterization of $\Gamma_i$, the choice $\Gamma_i, \text{true} = \theta_{\text{true}} \times 1_{d_i}1_{d_i}^T$ allows us to compare to the random intercept GLMM fit using MixedModels.jl. When the random effect term is a scalar, MixedModels.jl uses Gaussian quadrature for parameter estimation. We compare estimates and run-times to the random intercept GLMM fit of MixedModels.jl with 25 Gaussian quadrature points. We conduct simulation studies under two scenarios (simulation I and II). In simulation I, it is assumed that the data are generated by the quasi-copula model with $\theta_{\text{true}} = 0.1$, and in simulation II, it is assumed that the true distribution is the random intercept GLMM with $\theta_{\text{true}} = 0.01, 0.05$.

Figures 1-4 summarize the performance of the MLEs using mean squared errors (MSE) under the AR(1) parameterization of $\Gamma_i$. Figures 5-8 summarize the same under the CS parameterization of $\Gamma_i$. Figures 9-10 help us assess estimation accuracy and how well the GLMM density approximates the quasi-copula density under simulation I for the Bernoulli and Gaussian base. Under simulation II, Figures 11-14 shed light on how well the quasi-copula density approximates the GLMM density under different magnitudes of variance components. Figure 11 shows that for the Bernoulli base distribution with $\theta_{\text{true}} = 0.05$, the quasi-copula estimates of the variance component have average MSE of about $10^{-3}$. Figure 12 shows that for the Bernoulli base distribution with $\theta_{\text{true}} = 0.01$, the quasi-copula estimates of the variance component improves to an average MSE of about $10^{-4}$. Figure 13 shows for the Gaussian base distribution with $\theta_{\text{true}} = 0.05$, the quasi-copula estimates of the variance component has an average MSE around $10^{-4}$, and the quasi-copula estimates of the precision has an average MSE around $10^{-2}$. Figure 14 shows that for the Gaussian base distribution with $\theta_{\text{true}} = 0.01$, the quasi-copula estimates of the variance component improves to an average MSE of about $10^{-6}$, and the quasi-copula estimates of the precision improves to an average MSE of about $10^{-4}$. The QC model accurately estimates the mean components even with large cluster sizes ($d_i = 25$) and small sample sizes ($n = 100$), even when the true density is that of the GLMM and LMM.
11.5.1 AR(1) Covariance

Figure 8: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the AR(1) covariance for the Poisson base distribution with log link function. Each scenario reports involves 100 replicates.
Figure 9: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the AR(1) covariance for the negative binomial base distribution with log link function. Each scenario involves 100 replicates.
Figure 10: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the AR(1) covariance for the Bernoulli base distribution with logit link function. Each scenario reports involves 100 replicates.
Figure 11: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the AR(1) covariance for the Normal base distribution with Identity link function. Each scenario involves 100 replicates.
11.5.2 CS Covariance

Figure 12: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the CS covariance for the Poisson base distribution with log link function. Each scenario reports involves 100 replicates.
Figure 13: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the CS covariance for the negative binomial base distribution with log link function. Each scenario reports involves 100 replicates.
Figure 14: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the CS covariance for the Bernoulli base distribution with logit link function. Each scenario reports involves 100 replicates.
Figure 15: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the CS covariance for the Normal base distribution with Identity link function. Each scenario reports involves 100 replicates.
11.5.3 VC Covariance

Figure 16: Simulation I: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the Bernoulli base distribution with logit link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario reports involves 100 replicates.
Figure 17: Simulation I: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta$ under the Normal base distribution with Identity link function and a single VC versus a random intercept LMM fit via MixedModels.jl. Each scenario reports involves 100 replicates.
Figure 18: Simulation II: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta = 0.05$ under the Bernoulli base distribution with logit link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario reports involves 100 replicates.
Figure 19: Simulation II: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta = 0.01$ under the Bernoulli base distribution with logit link function and a single VC versus a random intercept GLMM fit via MixedModels.jl. Each scenario reports involves 100 replicates.
Figure 20: Simulation II: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta = 0.05$ under the Normal base distribution with Identity link function and a single VC versus a random intercept LMM fit via MixedModels.jl. Each scenario reports involves 100 replicates.
Figure 21: Simulation II: Mean squared errors (MSE) of parameter estimates $\beta$ and $\theta = 0.01$ under the Normal base distribution with Identity link function and a single VC versus a random intercept LMM fit via MixedModels.jl. Each scenario reports involves 100 replicates.

11.5.4 Run Times

Run times under simulation I and II are comparable. Tables 1-4 presents average run times and their standard errors in seconds, for 100 replicates under the AR(1) and CS covariance structures. Tables 5-6 present average run times and their standard errors in seconds, for 100 replicates under simulation II with $\theta_{\text{true}} = 0.01$. All computer runs were performed on a standard 2.3 GHz Intel i9 CPU with 8 cores. Runtimes for the quasi-copula model are presented using multi-threading across 8 cores.
| n   | d_i | Poisson AR(1) time | Poisson CS time |
|-----|-----|--------------------|-----------------|
| 100 | 2   | 0.057 ± 0.001     | 0.065 ± 0.002   |
| 100 | 5   | 0.069 ± 0.002     | 0.079 ± 0.002   |
| 100 | 10  | 0.112 ± 0.008     | 0.133 ± 0.011   |
| 100 | 15  | 0.214 ± 0.021     | 0.212 ± 0.019   |
| 100 | 20  | 0.238 ± 0.023     | 0.234 ± 0.020   |
| 100 | 25  | 0.307 ± 0.025     | 0.289 ± 0.023   |
| 1000 | 2   | 0.060 ± 0.001     | 0.066 ± 0.001   |
| 1000 | 5   | 0.074 ± 0.001     | 0.081 ± 0.001   |
| 1000 | 10  | 0.096 ± 0.001     | 0.108 ± 0.002   |
| 1000 | 15  | 0.112 ± 0.002     | 0.125 ± 0.002   |
| 1000 | 20  | 0.153 ± 0.012     | 0.158 ± 0.008   |
| 1000 | 25  | 0.153 ± 0.003     | 0.180 ± 0.011   |
| 10000 | 2  | 0.201 ± 0.002     | 0.199 ± 0.002   |
| 10000 | 5  | 0.271 ± 0.002     | 0.302 ± 0.003   |
| 10000 | 10 | 0.358 ± 0.002     | 0.446 ± 0.004   |
| 10000 | 15 | 0.447 ± 0.004     | 0.564 ± 0.006   |
| 10000 | 20 | 0.543 ± 0.005     | 0.651 ± 0.006   |
| 10000 | 25 | 0.703 ± 0.008     | 0.757 ± 0.007   |

Table 5: Run times and (standard error of run times) in seconds based on 100 replicates for Poisson Base under AR(1) and CS covariance structure with sampling unit size $d_i$ and sample size $n$. 
| n   | d_i | NB AR(1) time | NB CS time |
|-----|-----|---------------|------------|
| 100 | 2   | 0.323 (0.009) | 0.300 (0.009) |
| 100 | 5   | 0.339 (0.007) | 0.311 (0.008) |
| 100 | 10  | 0.320 (0.008) | 0.337 (0.012) |
| 100 | 15  | 0.334 (0.011) | 0.391 (0.016) |
| 100 | 20  | 0.364 (0.013) | 0.372 (0.015) |
| 100 | 25  | 0.376 (0.016) | 0.362 (0.016) |
| 1000| 2   | 0.445 (0.004) | 0.381 (0.004) |
| 1000| 5   | 0.499 (0.003) | 0.429 (0.004) |
| 1000| 10  | 0.564 (0.004) | 0.520 (0.009) |
| 1000| 15  | 0.654 (0.010) | 0.700 (0.021) |
| 1000| 20  | 0.798 (0.019) | 0.864 (0.030) |
| 1000| 25  | 0.938 (0.022) | 0.864 (0.030) |
| 10000|2   | 2.656 (0.012) | 2.297 (0.017) |
| 10000|5   | 3.161 (0.013) | 2.706 (0.012) |
| 10000|10  | 3.875 (0.015) | 4.001 (0.059) |
| 10000|15  | 4.924 (0.016) | 5.302 (0.140) |
| 10000|20  | 6.353 (0.028) | 6.073 (0.142) |
| 10000|25  | 7.449 (0.109) | 6.987 (0.144) |

Table 6: Run times and (standard error of run times) in seconds based on 100 replicates for negative binomial (NB) Base under AR(1) and CS covariance structure with sampling unit size $d_i$ and sample size $n$. 
| n   | d_i | Bernoulli AR(1) time | Bernoulli CS time |
|-----|-----|-----------------------|-------------------|
| 100 | 2   | 0.052 (0.002)         | 0.051 (0.002)     |
| 100 | 5   | 0.062 (0.002)         | 0.069 (0.003)     |
| 100 | 10  | 0.176 (0.019)         | 0.123 (0.012)     |
| 100 | 15  | 0.218 (0.021)         | 0.213 (0.017)     |
| 100 | 20  | 0.253 (0.022)         | 0.310 (0.021)     |
| 100 | 25  | 0.299 (0.024)         | 0.339 (0.021)     |
| 1000| 2   | 0.080 (0.002)         | 0.056 (0.002)     |
| 1000| 5   | 0.081 (0.001)         | 0.069 (0.002)     |
| 1000| 10  | 0.096 (0.006)         | 0.088 (0.001)     |
| 1000| 15  | 0.121 (0.006)         | 0.119 (0.007)     |
| 1000| 20  | 0.179 (0.016)         | 0.179 (0.015)     |
| 1000| 25  | 0.226 (0.020)         | 0.232 (0.020)     |
| 10000|2   | 0.183 (0.002)         | 0.171 (0.003)     |
| 10000|5   | 0.256 (0.002)         | 0.264 (0.003)     |
| 10000|10  | 0.304 (0.002)         | 0.356 (0.003)     |
| 10000|15  | 0.432 (0.004)         | 0.450 (0.004)     |
| 10000|20  | 0.507 (0.005)         | 0.535 (0.007)     |
| 10000|25  | 0.614 (0.005)         | 0.673 (0.007)     |

Table 7: Run times and (standard error of run times) in seconds based on 100 replicates for Bernoulli Base under AR(1) and CS covariance structure with sampling unit size $d_i$ and sample size $n$. 
| n   | d_i | Gaussian AR(1) time | Gaussian CS time |
|-----|-----|---------------------|------------------|
| 100 | 2   | 0.213 (0.008)       | 0.214 (0.008)    |
| 100 | 5   | 0.305 (0.021)       | 0.338 (0.022)    |
| 100 | 10  | 0.392 (0.025)       | 0.432 (0.027)    |
| 100 | 15  | 0.507 (0.028)       | 0.441 (0.029)    |
| 100 | 20  | 0.533 (0.027)       | 0.448 (0.031)    |
| 100 | 25  | 0.590 (0.027)       | 0.429 (0.030)    |
| 1000| 2   | 0.236 (0.006)       | 0.236 (0.006)    |
| 1000| 5   | 0.272 (0.005)       | 0.309 (0.006)    |
| 1000| 10  | 0.365 (0.011)       | 0.415 (0.010)    |
| 1000| 15  | 0.461 (0.024)       | 0.547 (0.021)    |
| 1000| 20  | 0.548 (0.028)       | 0.628 (0.026)    |
| 1000| 25  | 0.561 (0.030)       | 0.669 (0.026)    |
| 10000| 2  | 0.604 (0.013)       | 0.582 (0.011)    |
| 10000| 5  | 0.753 (0.016)       | 0.793 (0.017)    |
| 10000| 10 | 0.871 (0.015)       | 1.053 (0.015)    |
| 10000| 15 | 1.032 (0.018)       | 1.300 (0.022)    |
| 10000| 20 | 1.233 (0.030)       | 1.718 (0.025)    |
| 10000| 25 | 1.437 (0.033)       | 2.191 (0.042)    |

Table 8: Run times and (standard error of run times) in seconds based on 100 replicates for Gaussian Base under AR(1) and CS covariance structure with sampling unit size $d_i$ and sample size $n$. 

77
| n    | d_i | Bernoulli QC time | Bernoulli GLMM time |
|------|-----|-------------------|---------------------|
| 100  | 2   | 0.048 (<0.001)   | 0.022 (0.002)       |
| 100  | 5   | 0.049 (0.001)    | 0.041 (0.001)       |
| 100  | 10  | 0.050 (0.001)    | 0.086 (0.004)       |
| 100  | 15  | 0.049 (0.001)    | 0.125 (0.005)       |
| 100  | 20  | 0.047 (0.001)    | 0.167 (0.005)       |
| 100  | 25  | 0.047 (0.001)    | 0.203 (0.008)       |
| 1000 | 2   | 0.045 (0.001)    | 0.166 (0.003)       |
| 1000 | 5   | 0.045 (0.001)    | 0.446 (0.013)       |
| 1000 | 10  | 0.043 (0.001)    | 0.899 (0.022)       |
| 1000 | 15  | 0.044 (0.001)    | 1.435 (0.038)       |
| 1000 | 20  | 0.054 (0.002)    | 1.888 (0.041)       |
| 1000 | 25  | 0.077 (0.002)    | 2.461 (0.057)       |
| 10000| 2   | 0.138 (0.003)    | 1.726 (0.034)       |
| 10000| 5   | 0.160 (0.003)    | 4.711 (0.099)       |
| 10000| 10  | 0.189 (0.003)    | 10.389 (0.221)      |
| 10000| 15  | 0.232 (0.003)    | 15.958 (0.327)      |
| 10000| 20  | 0.276 (0.003)    | 21.609 (0.313)      |
| 10000| 25  | 0.349 (0.003)    | 28.723 (0.494)      |

Table 9: Run times and (standard error of run times) in seconds based on 100 replicates under simulation II with Bernoulli Base, \( \theta_{\text{true}} = 0.01 \), sampling unit size \( d_i \) and sample size \( n \).
Table 10: Run times and (standard error of run times) in seconds based on 100 replicates under simulation II with Gaussian Base, $\theta_{true} = 0.01$, sampling unit size $d_i$ and sample size $n$. 

| $n$  | $d_i$ | Gaussian QC time  | LMM time  |
|------|-------|-------------------|-----------|
| 100  | 2     | 0.112 (0.002)     | 0.003 (0.003) |
| 100  | 5     | 0.106 (0.003)     | 0.001 (<0.001) |
| 100  | 10    | 0.097 (0.002)     | 0.001 (<0.001) |
| 100  | 15    | 0.099 (0.004)     | 0.001 (<0.001) |
| 100  | 20    | 0.105 (0.008)     | 0.001 (<0.001) |
| 100  | 25    | 0.109 (0.008)     | 0.003 (0.002) |
| 1000 | 2     | 0.110 (0.002)     | 0.004 (0.002) |
| 1000 | 5     | 0.103 (0.002)     | 0.002 (<0.001) |
| 1000 | 10    | 0.100 (0.002)     | 0.006 (0.002) |
| 1000 | 15    | 0.095 (0.001)     | 0.006 (0.001) |
| 1000 | 20    | 0.094 (0.002)     | 0.008 (0.001) |
| 1000 | 25    | 0.099 (0.002)     | 0.011 (0.002) |
| 10000| 2     | 0.200 (0.005)     | 0.018 (0.003) |
| 10000| 5     | 0.192 (0.004)     | 0.029 (0.003) |
| 10000| 10    | 0.216 (0.004)     | 0.050 (0.004) |
| 10000| 15    | 0.219 (0.003)     | 0.067 (0.003) |
| 10000| 20    | 0.239 (0.003)     | 0.091 (0.003) |
| 10000| 25    | 0.258 (0.002)     | 0.099 (0.003) |