ISYE 6644: SIMULATION: EVALUATING CONVERGENCE OF HERITABILTIY ESTIMATES ACROSS TRAITS OF INTEREST VIA GENETIC MODELS AND SIMULATION

[redacted]

Georgia Institute of Technology, USA

ABSTRACT

Assessing the genetic impact of an individual is of crucial importance across a variety of domains in human health, disease, and function. Inferring this impact necessitates the utilization of twin data and genetic modelling techniques. Such genetic models make inherent assumptions about the data, and it is of interest to investigate the model performance across different conditions using simulation. For two models, Normal ACE model and Falconer's method, input data simulated from bivariate Lagrangian Poisson (blgp), multivariate t, and bivariate normal and relevant modulations. It was found that an increase in sample size resulted in a decrease in the average and true standard error of genetic estimates. When the data was normally distributed, the standard error (average and true) converged. The greater the normality of the simulated twin data, the greater performance both models had across coverage rates and average and true standard errors. When the model assumptions for ACE were violated (non-normality of data, unequal variances across MZ and DZ groups) the model resulted in biased estimates of heritability and associated coverage rates. Lastly, the coverage rates for Falconer's model appears to contradict that found in [1], but otherwise the results strongly converge to that of [1].

Index Terms— Twin study, heritability, simulation

1. BACKGROUND & DESCRIPTION OF PROBLEM

Assessing the enviornmental and genetic impact of an individual is of crucial importance across a variety of domains in human health, disease, and function. Inferring these factors necessitates either strict genetic data (one's genome) or, more often, the utilization of twin data [1]. Twins can be separated into two types, monozygotic (MZ) or dizygotic (DZ) [1]. What defines a twin in general is that they are born at the same time but in which two distinct processes had occurred even so. Specifically, MZ twins share 100% of their genomic information with one another, and DZ twins share 50% on average [1]. The similarity in MZ twins, and the fact that they were born at the same time point (essentially), allows for any differences in the twins to be more indicative of environmental influences than that of genetic influences (since, by definition, their genomes are identical). However, this attribution of genetic influences can also be decomposed further into additive genetics and dominance (the second of which is not an interest in this study) [1]. Specifically, it is possible, using both MZ and DZ twins to evaluate *heritability* which is a function of the genetic components (defined differently per model) of the trait of interest. Two models widely used in the twin study literature (and genetics literature) are Falconer's Formula and the Normal ACE (NACE) model, both of which make implict assumptions about the input data from MZ and DZ twins, and both of which have their own benefits and disadvantages. Hence, it is important to talk about the mathematical properties of both models.

The NACE model utilizes MZ and DZ twin data in order to model the heritability of a given trait of interest. In doing so, it effectively breaks down the trait covariance of each MZ or DZ twin into additive genetics (A), common shared environment (C), and nonshared environment (E) variance components [1]. This approach is done via structural equation modelling (SEM). For the NACE model, notably, there are key assumptions which allow it to achieve the results it does. Firstly, any given trait of interest (which is what would be investigated using the data of the MZ and DZ twin pairs that one has access to) is normally distributed [1]. Secondly, the NACE model assumes that the ACE variance parameters are equal for both MZ twin pairs and DZ twin pairs [1].

In contrast, another model, Falconer's Formula, which is a distribution free method of moment estimators [1] makes no assumptions regarding the variance across MZ twin pairs and DZ twin pairs (i.e. they are allowed to be unequal), with the additional assumption that the proportion of the total variance (for genetics or environmental effects) are the same [1].

Therefore there are key differences in assumptions between the NACE model and Falconer's formula. Furthermore, given the wide use of these models across the literature in looking at differently distributed traits (which are of wide interest in a variety of fields as they relate to genetics, human behavior, disease, and biomarkers), there is a great need to evaluate the effect of these assumptions that these models hold on the resultant parameter estimates in these scenarios. In doing so, this work can be considered for future traits of interest when utilizing models, the approach can be considered for other models of interest, and modulations therein can be expanded upon based on the implementation herein [1].

This paper subequently seeks to use the methods in [1] in the same manner to see how simulation applies in the context of evaluating these models' heritability estimates, across NACE model and Falconer's formula. By validating the findings in the paper, this will be an important step forward in looking at additional models of interest, and understanding more in-depth the process in which models can be assessed (whether it be for heritability or other metrics as well).

The upcoming sections of this paper are outlined as follows: Section 2) the mathematical theory of each model (NACE and Falconer), relevant assumptions and mathematical setup of how the analysis will be performed; section 3) discussion of the application, or general framework of how to do the simulation analysis and associated simulation concepts. Afterwards, in section 4), resulting visual demonstrations of the simulation will be shown for the various steps outlined in 3), and the results specific to the model comparisons will be demonstrated. Finally in 4), there will be a discussion of the results, and additionally whether it validates the paper of interest, and 5) the conclusion will state the overall takeaways along with planned future work relevant to the results. In brief, the techniques that will be used to investigate this question are a) input analysis (determining optimal distributions of interest for the traits selected based on literature review), b) simulation of data based on the input data analysis distribution, c) using both NACE and Falconer's methods on the simulated data, and d) output analysis using confidence intervals on the resultant model parameter estimates.

2. THEORY

2.1. Models

Following mostly from [1] (unless stated otherwise) the following equations relevant to models of interest (NACE and Falconer's) can be derived. From here on, the explanation will be from understanding or from the paper [1], and any derivations by hand will be noted. First, define the total amount of twin pairs for MZ twins as N_{MZ} and that of DZ twins as N_{DZ}, for a total number of twin pairs as $N = N_{MZ} + N_{DZ}$. For any given zygosity (MZ or DZ), z, define $\boldsymbol{y}_z = (y_{z_1}, y_{z_2})$, where $\boldsymbol{y}_z$ is the trait or phenotype of interest of the respective twin pair for a given zygoosity. Notably we can decompose this $\boldsymbol{y}_z$ then as $\boldsymbol{y}_z = \boldsymbol{x}_z^T\boldsymbol{\beta} + \boldsymbol{A}_z + \boldsymbol{C}_z + \boldsymbol{E}_z$, in which $\boldsymbol{x}_z^T$ is a $2 \times P$ matrtix of covariances for both twins (note we will expect that the expectation $\mathbb{E}(\boldsymbol{y}_z) = 0$ for all problems henceforth, i.e. that there will be no covariates). It then follows that $cov(\boldsymbol{y}_z) = \boldsymbol{\sigma}_z = cov(\boldsymbol{A}_z) + cov(\boldsymbol{C}_z) + cov(\boldsymbol{E}_z)$ and to create the model to have randomness, we define the parameters to be distributed as: $\boldsymbol{A}_z \sim (\boldsymbol{0}, \sigma^2_{A_z} K_z), \boldsymbol{C}_z \sim (\boldsymbol{0}, \sigma^2_{C_z} J), \boldsymbol{E}_z \sim (\boldsymbol{0}, \sigma^2_{E_z} I)$, where we define the matrices:

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, J = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, K_z = \begin{bmatrix} 1 & w_z \\ w_z & 1 \end{bmatrix}$$

Here is where w_z is defined as the 'genome relationship matrix' where $w_z = 1$ for MZ twins as they share 100% of their genes, and $w_z = 0.5$ for DZ twins since they share 50% of their genes. Based on these equations, then the overall variation of a phenotype can be decomposed into $\sigma^2_{A_z}, \sigma^2_{C_z}, \sigma^2_{E_z}$, which represents the additive genetic, shared environment, and unshared environment variances for twin of type zygosity z. Then, using these parameters as a baseline, we can estimate the heritability, an indication of the level of genetic importance of a trait as (the proportion of total trait variance due to additive genetics):

$$h^2 = \frac{\sigma^2_{A_{MZ}}}{\sigma^2_{A_{MZ}} + \sigma^2_{C_{MZ}} + \sigma^2_{E_{MZ}}} = \frac{\sigma^2_{A_{DZ}}}{\sigma^2_{A_{DZ}} + \sigma^2_{C_{DZ}} + \sigma^2_{E_{DZ}}}$$

Relatedly, the shared environment estimate (that is, the proportion of trait variance due to shared environmental effects) can be derived as:

$$c^2 = \frac{\sigma^2_{C_{MZ}}}{\sigma^2_{A_{MZ}} + \sigma^2_{C_{MZ}} + \sigma^2_{E_{MZ}}} = \frac{\sigma^2_{C_{DZ}}}{\sigma^2_{A_{DZ}} + \sigma^2_{C_{DZ}} + \sigma^2_{E_{DZ}}}$$

These parameters (along with $e^2 = (1 - h^2 - c^2)$ can be derived using NACE and Falconer's formula based on their inherent assumptions.
Then the multivariate normal distribution is given as (not from paper):

$$\mathbf{y_z} = (\mathbf{y_{z_1}}, \mathbf{y_{z_2}}) \sim \mathbf{N_2}(\mu = \mathbf{0}, \mathbf{\Sigma_z})$$
$$= f(\mathbf{y_z}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\mathbf{d/2}}\sqrt{\det \mathbf{\Sigma_z}}} \exp\Big(-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{y_z^T \Sigma_z^{-1} y_z}\Big)$$

Where d is the number of dimensions (in this case $d = 2$, so the exponent for the 2π term goes to 1.). We can then find the log likelihood (derived by hand) such that:

$$\log(f(\boldsymbol{y}_z|\boldsymbol{\alpha})) = \log(1) - \log(2\pi) - \frac{1}{2}\log(\det \Sigma_z) - \frac{1}{2}\mathbf{y_z^T \Sigma_z^{-1} y_z}$$
$$\log(f(\boldsymbol{y}_z|\boldsymbol{\alpha})) = -\frac{1}{2}\log(\det \Sigma_z) - \frac{1}{2}\boldsymbol{y}_z^T\Sigma_z^{-1}\mathbf{y_z} - \log(\mathbf{2}\pi)$$
$$\log(f(\boldsymbol{y}_z|\boldsymbol{\alpha})) = -\frac{1}{2}\log(|\Sigma_z|) - \frac{1}{2}\boldsymbol{y}_z^T\Sigma_z^{-1}\mathbf{y_z} - \log(\mathbf{2}\pi)$$
$$\log(f(\boldsymbol{y}_z|\boldsymbol{\alpha})) = -0.5(\log(|\mathbf{\Sigma}_z| + \boldsymbol{y}_z^T\Sigma_z^{-1}\boldsymbol{y}_z + 2\log(2\pi))$$

where $\mathbf{\Sigma}_z = \begin{bmatrix} \sigma^2_A + \sigma^2_C + \sigma^2_E & w_z\sigma^2_A + \sigma^2_C \\ w_z\sigma^2_A + \sigma^2_C & \sigma^2_A + \sigma^2_C + \sigma^2_E \end{bmatrix}$, and $\boldsymbol{\alpha} = (\sigma^2_A, \sigma^2_C, \sigma^2_E)$. Further, the NACE model then assumes that:

$\sigma^2_{A_z} = \sigma^2_A, \sigma^2_{C_z} = \sigma^2_C, \sigma^2_{E_z} = \sigma^2_E$, and ther covariance terms can be re-represented in terms of the previously defined matrices, I, J, and K to re-represent the covariance matrix: $cov(\boldsymbol{y}_z) = \boldsymbol{\Sigma}_z = \sigma^2_A \boldsymbol{K}_z + \sigma^2_C \boldsymbol{J} + \sigma^2_E \boldsymbol{I}$. We can then derive the estimating equations for $\boldsymbol{\alpha}$ by taking the derivative with respect to zero of the log likelihood function and setting it to 0 to find where there exists local optima:

$$\boldsymbol{u}(\boldsymbol{\alpha})_{NACE} = \frac{\partial}{\partial \alpha} logf(\boldsymbol{y}_z|\boldsymbol{\alpha})$$
$$= \Big(\frac{\partial}{\partial \sigma^2_A} logf, \frac{\partial}{\partial \sigma^2_C} logf, \frac{\partial}{\partial \sigma^2_E} logf\Big)^T = \mathbf{0}$$

We can do this because the log function is a one-to-one function, meaning that even taking the log and then the derivative will result in the same outcome as just taking the derivative (it is just simpler to find the maximum likelihood estimation through the log-likelihood). This result will obtain our $\hat{\alpha}$ estimates. Furthermore, based on the preceding definitions, and under the regularity constraints of maximum likelihood estimation, we can derive the Fisher-Information matrix (the expectation of the second derivative of the log likelihood function) [3]:

$$\boldsymbol{V} = -\mathbb{E}\Big(\frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T}\Big)$$

Notably by the properties of the maximum likelihood estimation (MLE) from [4] we have that for the given Fisher Information matrix above:

$$\sqrt{N}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \rightarrow^D MVN(\mathbf{0}, \boldsymbol{V^{-1}})$$

And this corroborates with the multivariate central limit theorem which states that:

$$\sqrt{n}(\bar{\boldsymbol{X}} - \mu) \rightarrow^D N_k(0, \boldsymbol{\Sigma})$$

Based on that logic in the case of α we know that $\boldsymbol{\Sigma} = V^{-1}$. Altogether, this implies that as based on MLE properties, that the asymptotic lower bound of the variance can be equated to the inverse of V, or the inverse of the Fisher Information matrix [3,4]. Subsequently we define the covariance as [1]:

$$\hat{cov}(\hat{\alpha}) = \frac{1}{N}\hat{V}^{-1} = \frac{1}{N}\Big[\frac{-1}{N}\sum_1^N \frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} logf(\boldsymbol{y}_z|\boldsymbol{\alpha})\Big]^{-1}_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}$$

And from these two estimates we therefore finalize the resultant $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{cov}}(\boldsymbol{\alpha})$ which can be used to construct the confidence intervals. Notably, when actually running the ACE model through the specific implementation in R, we derive the square root of the estimates along with the standard errors, which requires that we actually transform these estimates into the normal variance space. In order to do so, we must transform the standard errors using the delta method as derived (by hand) below for the way it was done in the package for both estimates that were calculated from the square root of the variance components (h^2 and c^2 from the σ components). First given a parameter specific delta method as given by (assuming square root of α):

$$\sqrt{n}(\sqrt{\boldsymbol{\alpha}} - \hat{\sqrt{\boldsymbol{\alpha}}}) \rightarrow^D N(0, \Sigma)$$
$$h(\hat{\sqrt{\boldsymbol{\alpha}}}) \approx h(\hat{\sqrt{\boldsymbol{\alpha}}}) + \nabla h(\hat{\sqrt{\boldsymbol{\alpha}}})^T(\sqrt{\boldsymbol{\alpha}} - \hat{\sqrt{\boldsymbol{\alpha}}})$$
$$var(h(\hat{\sqrt{\boldsymbol{\alpha}}})) \approx var(h(\hat{\sqrt{\boldsymbol{\alpha}}}) + \nabla h(\hat{\sqrt{\boldsymbol{\alpha}}})^T(\sqrt{\boldsymbol{\alpha}} - \hat{\sqrt{\boldsymbol{\alpha}}}))$$
$$= var(\nabla h(\hat{\sqrt{\boldsymbol{\alpha}}})^T \sqrt{\boldsymbol{\alpha}})$$
$$= \nabla h(\hat{\sqrt{\boldsymbol{\alpha}}})^T cov(\sqrt{\boldsymbol{\alpha}}) \nabla h(\hat{\sqrt{\boldsymbol{\alpha}}})$$
$$= \nabla h(\hat{\sqrt{\boldsymbol{\alpha}}})^T (\Sigma) \nabla h(\hat{\sqrt{\boldsymbol{\alpha}}})$$

where Σ in this context is the covariance matrix of the square root of the variance components output by the package used in this study.

Based on the delta method for the square root of alpha, we would use the following h functions to derive our heritability and shared environment estimates so as to transform the standard errors to this space (derived by hand):

$$h^2 = \frac{(\sqrt{\sigma^2_A})^2}{(\sqrt{\sigma^2_A})^2 + (\sqrt{\sigma^2_C})^2 + (\sqrt{\sigma^2_E})^2}$$
$$\frac{\partial h^2}{\partial \sqrt{\sigma^2_A}} =$$
$$\frac{((\sqrt{\sigma^2_A})^2 + (\sqrt{\sigma^2_C})^2 + (\sqrt{\sigma^2_E})^2)(2\sqrt{\sigma^2_A}) - (\sqrt{\sigma^2_A})^2(2\sqrt{\sigma^2_A})}{(\sqrt{\sigma^2_A})^2 + (\sqrt{\sigma^2_C})^2 + (\sqrt{\sigma^2_E})^2)^2}$$
$$= \frac{(\sigma^2_C + \sigma^2_E)(2\sqrt{\sigma^2_A})}{(\sigma^2_A + \sigma^2_C + \sigma^2_E)^2}$$
$$\frac{\partial h^2}{\partial \sqrt{\sigma^2_C}} = -\frac{\sigma^2_A(2\sqrt{\sigma^2_C})}{(\sigma^2_A + \sigma^2_C + \sigma^2_E)^2}$$
$$\frac{\partial h^2}{\partial \sqrt{\sigma^2_E}} = -\frac{\sigma^2_A(2\sqrt{\sigma^2_E})}{(\sigma^2_A + \sigma^2_C + \sigma^2_E)^2}$$

We then have the corresponding $h(\hat{\sqrt{\boldsymbol{\alpha}}})$ function for h^2:

$$\nabla h_h(\hat{\sqrt{\boldsymbol{\alpha}}}) = \Big(\frac{\partial h^2}{\partial \sqrt{\sigma^2_A}}, \frac{\partial h^2}{\partial \sqrt{\sigma^2_C}}, \frac{\partial h^2}{\partial \sqrt{\sigma^2_E}}\Big)$$

We can derive (by hand) an equaivalent expression for the estimate of c^2 and transforming the standard error of $\sqrt{\sigma^2_C}$ to

c^2 space.

$$\frac{\partial c^2}{\partial \sqrt{\sigma_C^2}} = -\frac{(\sigma_A^2+\sigma_E^2)(2\sqrt{\sigma_E^2})}{(\sigma_A^2+\sigma_C^2+\sigma_E^2)^2}$$
$$\frac{\partial c^2}{\partial \sqrt{\sigma_A^2}} = -\frac{\sigma_C^2(2\sqrt{\sigma_A^2})}{(\sigma_A^2+\sigma_C^2+\sigma_E^2)^2}$$
$$\frac{\partial c^2}{\partial \sqrt{\sigma_E^2}} = -\frac{\sigma_C^2(2\sqrt{\sigma_E^2})}{(\sigma_A^2+\sigma_C^2+\sigma_E^2)^2}$$
$$\nabla h_c(\hat{\sqrt{\boldsymbol{\alpha}}}) = \Big(\frac{\partial c^2}{\partial \sqrt{\sigma_A^2}}, \frac{\partial c^2}{\partial \sqrt{\sigma_C^2}}, \frac{\partial c^2}{\partial \sqrt{\sigma_E^2}}\Big)$$

From there, we can then compute the standard errors directly using the delta method as aforementioned. Once that is done, we can compute 95% confidence intervals using a z_{α_2} score equal to 1.96 ($\alpha = 0.05$), and then construct the confidence intervals such that:

$$h^2 \pm \boldsymbol{z}_{\alpha/2}(SE)$$
$$c^2 \pm \boldsymbol{z}_{\alpha/2}(SE)$$

where the standard error for either h^2 or c^2 is derived again by the delta method (once we have the variance $var(h(\hat{\sqrt{\boldsymbol{\alpha}}})$ as shown in the delta method we just have to take the square root and we would get our transformed standard error for both functions).

Then, similar to the NACE model, Falconer's formula is serived as follows, with the heritability h^2 and shared enviornment estimates c^2 as:

$$\hat{h}^2_{Falc} = 2(r_{MZ} - r_{DZ}), \hat{c}^2_{Falc} = 2r_{DZ} - r_{MZ}$$

where r_{MZ} is the pearson correlation within MZ twin pairs, and r_{DZ} is the pearson correlation within DZ twin pairs. Falconer's estimators are then derived [1] as:

$$\rho_{MZ} = Corr(\boldsymbol{y}_{MZ_1}, \boldsymbol{y}_{MZ_2}) = \frac{cov_{MZ}}{var(\boldsymbol{y}_{MZ})}$$
$$= \frac{\sigma^2_{A_{MZ}} + \sigma^2_{C_{MZ}}}{\sigma^2_{A_{MZ}} + \sigma^2_{C_{MZ}} + \sigma^2_{E_{MZ}}} = h^2 + c^2$$
$$\rho_{DZ} = Corr(\boldsymbol{y}_{DZ_1}, \boldsymbol{y}_{DZ_2}) = \frac{cov_{DZ}}{var(\boldsymbol{y}_{DZ})}$$
$$= \frac{\sigma^2_{0.5A_{DZ}} + \sigma^2_{C_{DZ}}}{\sigma^2_{A_{DZ}} + \sigma^2_{C_{DZ}} + \sigma^2_{E_{DZ}}} = 0.5h^2 + c^2$$

By definition from the previous defined equations Falconer's heritability and shared environment are then:

$$\Rightarrow 2(\rho_{MZ} - \rho_{DZ}) = h^2$$
$$2\rho_{DZ} - \rho_{MZ} = c^2$$

where ρ_{MZ} and ρ_{DZ} are the population correlation coefficients within MZ twins and within DZ twins respectively. One possible means of computing the standard error for Falconer's formulas is by considering the asymptotic variance of the pearson correlation coefficient which ultimately yields the following standard error for the two estimates of interest:

$$\hat{SE}(\hat{h}^2_{Falc}) \approx \sqrt{4(\hat{var}(r_{MZ}) + \hat{var}(r_{DZ})}$$
$$= \sqrt{4\Big(\frac{(1-r^2_{MZ})^2}{N_{MZ}} + \frac{(1-r^2_{DZ})^2}{N_{DZ}}\Big)}$$
$$\hat{SE}(\hat{c}^2_{Falc}) \approx \sqrt{4(\hat{var}(r_{DZ}) + \hat{var}(r_{MZ})}$$
$$= \sqrt{4\Big(\frac{(1-r^2_{DZ})^2}{N_{DZ}}\Big) + \frac{(1-r^2_{MZ})^2}{N_{MZ}}}$$

Once these are obtained in the study of interest, the confidence intervals for h^2 and c^2 will be constructed using a 95% confidence interval in the same way as NACE (except no need for the delta method). Therefore, it would be calculated again such that:

$$h^2 \pm \boldsymbol{z}_{\alpha/2}(SE)$$
$$c^2 \pm \boldsymbol{z}_{\alpha/2}(SE)$$

2.2. Distributions

Three distributions will be used in this study and they are derived as follows (no particular order).
First, the heavy weighted multivariate t-distribution is given as [1]:

$$\boldsymbol{y}_z = (y_{z_1}, y_{z_2}) \sim f(\boldsymbol{y}_z)$$
$$= \frac{\Gamma(\frac{\nu+2}{2})}{\Gamma(\frac{\nu}{2})\nu\pi|\Sigma_z|^{1/2}}\Big[1 + \frac{1}{\nu}\boldsymbol{y}_z^T\boldsymbol{\Sigma}_z^{-1}\boldsymbol{y}_z\Big]^{\frac{-(\nu+2)}{2}}$$
$$\boldsymbol{y}_z = (y_{z_1}, y_{z_2}) \sim bLGP(\sigma^2_A + \sigma^2_C + \sigma^2_E, \lambda)$$

Next, we can also define and sample discrete count variables from the bivariate Lagrangian Poisson (BLGP) [1] distribution eas follows for MZ twins:

$$Q_0 \sim LGP(\sigma^2_A + \sigma^2_C, \lambda)$$
$$Q_1, Q_2 \sim LGP(\sigma^2_E, \lambda)$$
$$Y_1 = Q_0 + Q_1$$
$$Y_2 = Q_0 + Q_2$$
$$\Rightarrow Y_1, Y_2 \sim bLGP(\sigma^2_A + \sigma^2_C + \sigma^2_E, \lambda)$$

and for DZ twins:

$$Q_0 \sim LGP(0.5\sigma_A^2 + \sigma_C^2, \lambda)$$
$$Q_1, Q_2 \sim LGP(0.5\sigma_A^2 + \sigma_E^2, \lambda)$$
$$Y_1 = Q_0 + Q_1$$
$$Y_2 = Q_0 + Q_2$$
$$\Rightarrow Y_1, Y_2 \sim bLGP(\sigma_A^2 + \sigma_C^2 + \sigma_E^2, \lambda)$$

Again, we can sample from the multivariate normal distribution (bivariate in this case, with d = 2) based on the predefined distribution above:

$$\mathbf{y_z} = (\mathbf{y_{z_1}}, \mathbf{y_{z_2}}) \sim \mathbf{N_2}(\mu = \mathbf{0}, \mathbf{\Sigma_z})$$
$$= f(\mathbf{y_z}) = \frac{\mathbf{1}}{(\mathbf{2}\pi)^{\mathbf{d/2}}\sqrt{\det \mathbf{\Sigma_z}}} \exp\Big(-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{y_z^T \Sigma_z^{-1} y_z}\Big)$$

2.3. Other

3. APPLICATION

First, for the distributions in theory mentioned, a single dataset distribution for 700 MZ and 700 DZ pairs was generated from each (normal, BLGP, t) to visualize what the samples look like (if they were expected, etc.). Specifically, for all distributions, the parameter estimates for the covariance matrix or as direct inputs were given as: $\sigma_A^2 = 0.5, \sigma_C^2 = 0.3$ and $\sigma_E^2 = 0.2$. For the t-distribution specificially, the degrees of freedom $\nu = 4.5$ was used, and for BLGP, the value of $\lambda = 0.35$ was used.
The sample normal distribution results was also tested using the Anderson-Darling test in order to determine normality (to assure that the distribution was multivariate normal).
Beyond this initial data visualizaton, three simulations were carried out based on the predefined theory. Using the same aforementioned distribution parameters, these samples were again sampled from for each simulation performed with the same parameters.

3.1. Simulation 1: Effect of distributions on parameter estimates and model

Using 1000 datasets (repetitions) of 700 MZ and DZ twin pairs each across the bivariate normal, t, and BLGP distributions, the input data to both the NACE and Falconer's methods were implemented and simulated. The average mean value $\bar{h^2}$ was calculated from the 1000 h^2 estimates per model, per distribution. The average standard error (derived from the according theory on delta method for NACE, and otherwise square root of the variance found in Falconer's), $\bar{SE}$ was found by averaging the resultant standard errors from each 1000 simulated estimates, per model, per distribution. The true standard error, denoted as SE in the resulting tables, was found as the square root of the variance across all 1000 simulated h^2 estimates), and the coverage rate (Cov) across all 95% confidence intervals generated (again through the previous delta method, z-values and SEs as in the theory section), was calculated. Specifically, for coverage, the true parameter estimate was determined as those given above $\sigma_A^2 = 0.5, \sigma_C^2 = 0.3$ since the resultant heritability and shared environment parameters are such that $c^2 = \sigma_C^2/(\sigma_A^2 + \sigma_C^2 + \sigma_E^2) = 0.3$, and $h^2 = \sigma_A^2/(\sigma_A^2 + \sigma_C^2 + \sigma_E^2)0.5$ by the same logic. Therefore, if the confidence intervals for a given model contained these true parameter values within the, these were included in the coverage rate (the amount of times that these true parameter values fell within the confidence interval upper and lower bounds). This was computed across each specific 1000 simulations (from the respective distributions), for a single coverage rate value per model of the estimate.

This exact same procedure was done for c^2. That is, for 1000 repetitions from a given distribution (normal, t, BLGP) of 700 MZ and DZ twins each simulation, for both models (NACE and Falconer's) respectively, the $\bar{c^2}, \bar{SE}, SE$ and Cov all assocaited with the estimate were determined and reported. The resuls for this simulation can be found in Table 1 and Table 2. This procedure was an extension of that from [1]. This specific simulation sought to see the effect of the distributions on the different parameter estimate outputs as well as any difference in the models with respect to these.

3.2. Simulation 2: Effect of sample size on heritability parameter estimate

Just as in the previous simulation, 1000 datasets (repetitions) were simulated, except in this case, across $N = N_{MZ} = N_{DZ} = 50, 100, 200, 400$ and 700. This was an extension of that from [1]. This simulation procedure was done only on the heritability estimates compared to the previous simulation. The same values as in simulation 1 were reported for each sample size: $\bar{h^2}, \bar{SE}, SE$ and Cov, all associated to the parameter estimate $\hat{h^2}$. The results from this simulation can be found in Tables 3,4 and 5 in the results section.

3.3. simulation 3: Effect of varying variance on heritability estimate for models

Similar to the previous simulations, again 1000 repetitions were used, but this time the focus was on varying the related variance of the MZ input data versus that of the DZ input data to see its impact on the assumptions of the models. That is,

$$var(y_{DZ}) = var(y_{DZ_2}) = \sigma_{A_{DZ}}^2 + \sigma_{C_{DZ}}^2 + \sigma_{E_{DZ}}^2$$
$$= .5 + .3 + .2 = 1$$
$$var(y_{MZ}) = \tau * var(y_{DZ}), \tau = [0.5, 0.6, 0.7, 0.8, 0.9, 1]$$

So for every τ, there was 700 MZ and 700 DZ twins simulated with 1000 repetitions. This was then plot as in Figure 6. The bias in the average heritability estimate bias for each 1000 bias for each iteration was calculated (see formula in Fig 6).

4. RESULTS

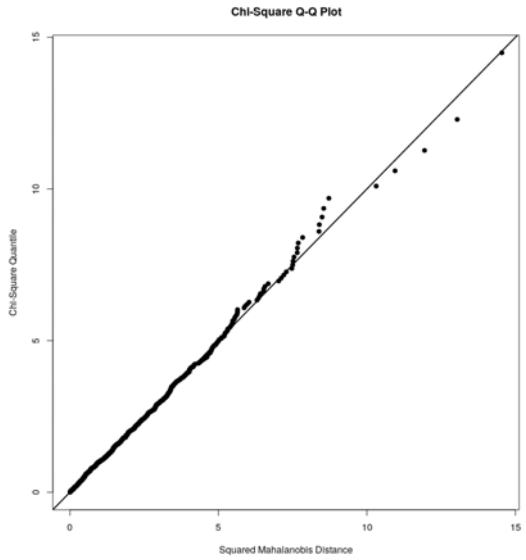


Fig. 1: QQ-plot and Anderson-Darling Test for multivariate normality across one of the MZ twin pair population samples generated from the bivariate normal distribution (for a given dataset, this would have 700 pairs). As can be observed, on the QQ-plot there is high convergence of the chi squared test statistic to the diagonal line, indicating normality. The overall AD test-statistic was: 0.2745643, and the resultant p-value: 0.8639136. A p-value > 0.15 indicates normality. Significance level is alpha = 0.05.

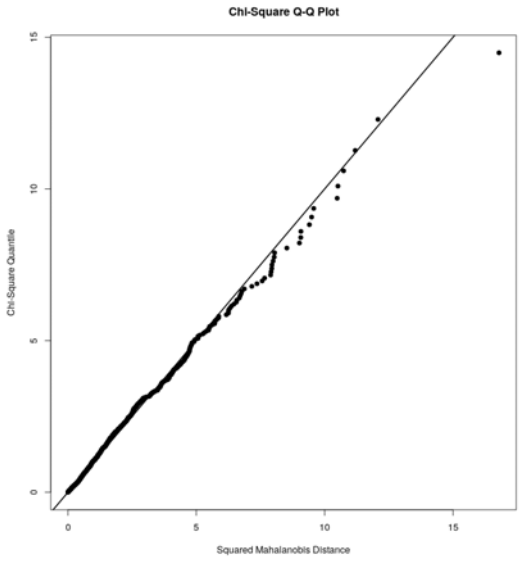


Fig. 2: QQ-plot and Anderson-Darling Test for multivariate normality across one of the DZ twin pair population samples generated from the bivariate normal distribution (for a given dataset, this would have 700 pairs). As can be observed, on the QQ-plot there is high convergence of the chi squared test statistic to the diagonal line, indicating normality. The overall AD test-statistic was: 0.707191, and the resultant p-value: 0.2727727. A p-value > 0.15 indicates normality. Significance level is alpha = 0.05.

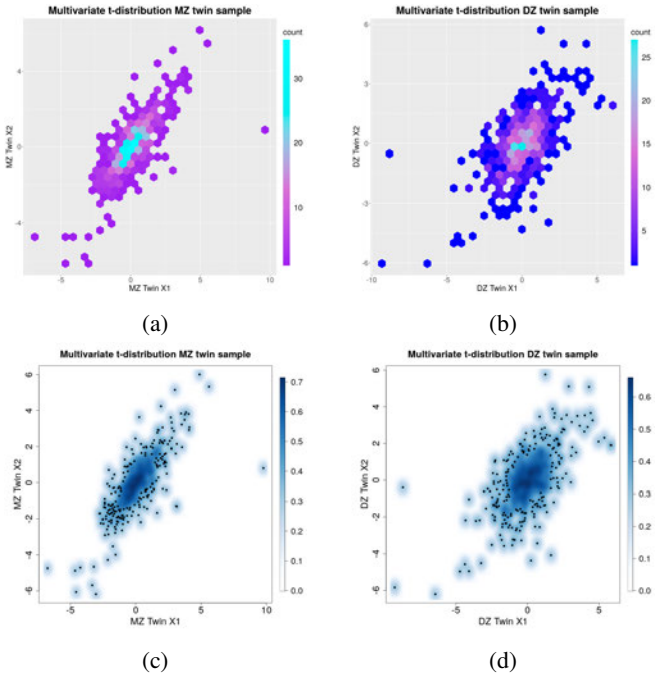
5. DISCUSSION

Overall, this study utilized simulation to robustly test the convergence in terms of performance of the NACE and Falconer's methods of genetic modelling of twins through a variety of modulations of parameters (sample size, variance equality, and distributions used).

The major findings pertain to the figures and tables that were output. In particular, Fig.1 demonstrated that as expected, twin pairs sampled from the bivariate normal distribution with the allocated covariance matrix based on the parameters chosen from [1] when tested with the Anderson-Darling test were found to be multivariate normal. Fig.2. demonstrates that this was also the case for the DZ twin pair population, albeit with a lower p-value (this could be because DZ twin pairs had a wider variance spread in both dimensions possible than that of MZ, but remains to be determined. Clearly, though, Anderson-Darling is a sufficient test to test for the multivariate normality of a distribution. Additionally, it confirms the expected theory of sampling and simulating from a normal distribution, so it attests to the package behaving as expected (albeit only a sample of one).

For Figures 3-5, the results are again as anticipated, with the main variation occuring based on whether the twin was MZ or DZ, and the spread (covariance) is clearly wider in DZ twins as would be expected based on the parameters initially set, and which is what is desired to simulate the theoretical understanding of DZ twins (vs MZ twins). For more discussion on the results of Fig3-5, there is add detail in the figure captions.

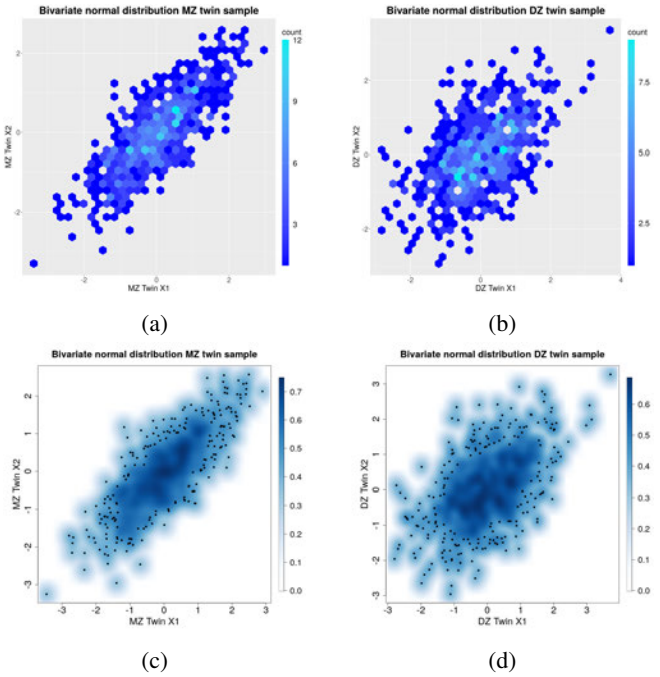
For Figure 6 (as explained in the figure caption), Falconer's method performed very well and in an unbiased fashion when considering the heritability estimate bias percentage as a function of differing MZ-DZ variance ratios. Conversely, the NACE model did not. As explained in [1] and which is a fundamental problem with the NACE model in the literature at large, this result matches with expected theory in terms of the disadvantages of the NACE model. Therefore, it is clear that when considering the twin samples, that the fundamental assumption of the variances across MZ and DZ twins being equivalent be considered carefully, otherwise, the model risks outputting a considerably biased resulting parameter estimate. Conversely, for Falconer's method, it only assumes the proportions of the variance components are equal, and not the actual variance components themselves,



(a) (b)

(c) (d)

Fig. 3: Heatmap of the resulting trait observations drawn from a/c) the multivariate t-distribution for MZ twins, and b/d) the multviariate t-distribution for DZ twins. Essentially, the distribution represents the actual multivariate t-distribution with the parameters input into the model as $\sigma_a^2 = 0.5, \sigma_c^2 = 0.3, \sigma_e^2 = 0.2, \nu = 4.5$ for the MZ covariance matrix and for the DZ covariance matrix.
As expected, for a/c) the t-distribution for MZ twin pairs sampled demonstrates a sharper correlation along one dimension than the other (i.e. less variance in one dimension). This result would be anticipated, because MZ twins should vary less on a given trait than DZ twins by definition. Conversely, the trait for DZ twins have a wider variance across both dimensions, which is what would be expected given their greater dissimilarity compared to MZ twins. Note the tails of the distribution are quite spread apart (as will be shown in contrast to the normal distribution in Figure 4). Note that the greatest density (or where the distribution is centered at) is toward the middle of the distribution, as would be expected. Further, the distributions are roughly symmetric.



(a) (b)

(c) (d)

Fig. 4: Heatmap of the resulting trait observations drawn from a/c) the bivariate normal distribution for MZ twins, and b/d) the bivariate normal distribution for DZ twins.
Essentially, the distribution represents the actual bivariate normal distribution with the parameters input into the model as $\sigma_a^2 = 0.5, \sigma_c^2 = 0.3, \sigma_e^2 = 0.2$ for the MZ covariance matrix and for the DZ covariance matrix. As expected, for a/c) the normal distribution for MZ twin pairs sampled demonstrates a sharper correlation along one dimension than the other (i.e. less variance in one dimension than the other). This result would be anticipated, because MZ twins should vary less on a given trait than DZ twins by definition. Conversely, the trait for DZ twins have a wider variance across both dimensions, which is what would be expected given their greater dissimilarity compared to MZ twins. Note the tails of the distribution are not spread out considerably (as in contrast to the heavy tailed t-distribution in Figure 3). Note that the greatest density (or where the distribution is centered at) is toward the middle of the distribution, as would be expected. Further, the distributions are roughly symmetric.

so it does not run into this issue as shown. This is a critical demonstration and successful replication of the paper that must not be understated.

For Table 1, it can be observed that regardless of the model or distribution, the $\bar{h}^2$ estimate was as expected, at a value 0.5 which is what the modelled σ_A^2 was initialized to (again our actual expected heritability parameter would be this additive variance divided by the total variances which already sum up to 1, and this means that the expected heritability estimate would be 0.5; so for the estimate to converge to this value demonstrates its success). Notably, the average standard error of the heritability estimate in general varied from the 'true' standard error estimate until the normal distribution was considered. It was notable that the performance (average SE vs true SE) became slightly better from the BLGP distribution to the t-distribution for both models, and then converged during the normal distribution. This logically follows from the NACE model, because its most inherent assumption is that the data are normally distributed. This fact is further exemplified by looking at the coverage across the

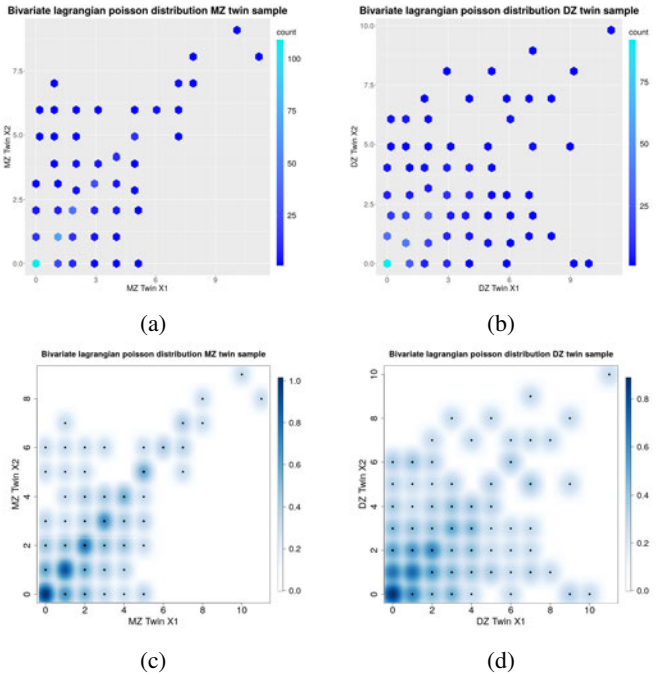


Fig. 5: Heatmap of the resulting trait observations drawn from a/c) the bivariate lagrangian poisson (blgp) distribution for MZ twins, and b/d) the blgp for DZ twins.

Essentially, the distribution represents the actual blgp distribution with the parameters input into the model as $\sigma_a^2 = 0.5, \sigma_c^2 = 0.3, \sigma_e^2 = 0.2, \lambda = 0.35$ for the MZ and DZ bivariate lagrangian poisson variance input. As expected, for a/c) the normal distribution for MZ twin pairs sampled demonstrates a sharper correlation along one dimension than the other (i.e. less variance in one dimension than the other). This result would be anticipated, because MZ twins should vary less on a given trait than DZ twins by definition. Conversely, the trait for DZ twins have a wider variance across both dimensions, which is what would be expected given their greater dissimilarity compared to MZ twins. Note that these are discrete outcomes as that is what the Poisson distribution is concerned with (as opposed to the normal and t-distributions earlier). Furthermore, note that the greatest density is at around (0,0), which is what would be expected based on the Poisson distribution (for both MZ and DZ twins).

BLGP, t and normal distributions in that order. The coverage increases as the distribution becomes more normal, in essence, and the coverage is almost 100% when a normal distribution is being simulated. Therefore, it is clear that the traits being investigated, when it comes to MZ or DZ twins, must be as normal as possible, otherwise there is great risk in lowering the coverage of the estimate, for the NACE model, but also based on the results, Falconer's formula as well.

Notably Table 2 demonstrates the same kind of results except in the context of c^2, the shared environment parameter

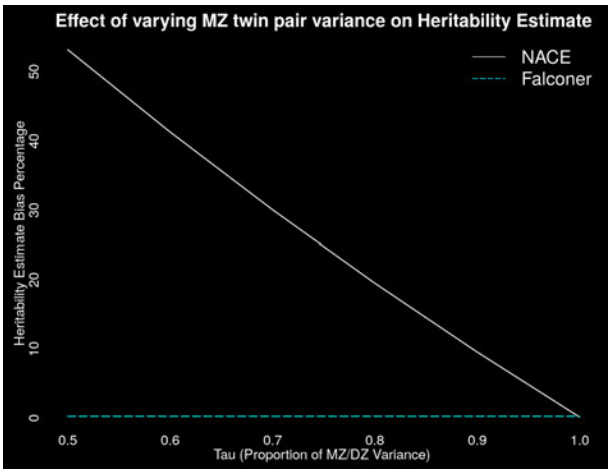


Fig. 6: Effect of varying the MZ twin pairs' variance as a function of DZ twin pairs' variance on the heritability estimate $\hat{h}^2$ across NACE and Falconer's model. 1000 datasets of 700 MZ and 700 DZ pairs were generated from a bivaraite normal distribution as prior, and input into each model, with the modulation that the variances differed between MZ and DZ twins. Specifically, Tau = $\tau = Var(y_{mz})/Var(y_{dz})$, and tau varied from 0.5 to 1 with 0.1 increments for this dataset. As is seen, given that a core assumption of the NACE model is that $\sigma_{MZ} = \sigma_{DZ}$, for all respective variance components (genetic, shared environment, unshared environment), when this assumption is violated, there is not consistent convergence of the heritability estimate $\hat{h}^2$ until $\tau = 1$, as compared to Falconer's Formula which is consistently low biased (or unbiased). The heritability bias was estimated as $(\hat{h}^2 - h^2)/h^2 * 100\%$ Falconer's formula makes no such assumption about the equality of variances, and hence does not have this issue.

estimate. As was the case for the average h^2 in Table 1, the average shared environment estimate matched with the expected 0.3, in roughly all the cases. The coverage, yet again, regardless of the model, was highest for the normally distributed data.

In terms of the findings for Table 3, by observing the table, it was found that as the number of pairs increased for the input into either model when performing the same type of simulation as in the previous contexts (except only for the heritability estimate), the average standard error and the 'true' standard error decrease. That result is expected because the standard error is a function of the variance, and if the sample size is increasing, then the variance in the result should seek to decrease by the central limit theorem. By definition, the standard error is a function of $\frac{\sigma}{\sqrt{n}}$, and with higher sample size, this should inevitably decrease (and ideally converge to a stable estimate). However, in this case, the average standard error and the 'true' standard error did not converge, for this BLGP distribution, which could speak more to the fact

Model	h^2	SE	SE	Cov
ACE-blgp	0.50	0.05	0.12	0.63
Falc-blgp	0.51	0.06	0.12	0.67
ACE-t	0.50	0.05	0.10	0.73
Falc-t	0.50	0.06	0.11	0.74
ACE-norm	0.50	0.05	0.05	0.96
Falc-norm	0.50	0.06	0.06	0.95

Table 1: h^2 estimates for NACE and Falconer's Formula across BLGP, t and Normal (norm) Distributions in 1000 simulations of 700 MZ and 700 DZ twin pairs from these distributions.
(ACE = NACE, Falc = Falconer's Formula, blgp = bivariate lagrangian poisson distributed data, t indicates multivariate t-distributed data, $\bar{h}^2$ = average estimate of h^2 after 1000 simulations, $\bar{SE}$ =the average of the h^2-associated standard errors for all 1000 simulations, SE the true standard error of all the 1000 estimates of h^2, and Cov = coverage of the true parameter σ_a^2 within the 1000 the 95% constructed confidence intervals across the datasets.)

Model	$\bar{c}^2$	SE	SE	Cov
ACE-blgp	0.29	0.05	0.10	0.66
Falc-blgp	0.29	0.05	0.10	0.69
ACE-t	0.30	0.05	0.09	0.75
Falc-t	0.30	0.05	0.10	0.75
ACE-norm	0.30	0.05	0.05	0.95
Falc-norm	0.30	0.05	0.05	0.95

Table 2: c^2 estimates for NACE and Falconer's Formula across BLGP, t and Normal (norm) Distributions in 1000 simulations of 700 MZ and 700 DZ twin pairs from these distributions.
(ACE = NACE, Falc = Falconer's Formula, blgp = bivariate lagrangian poisson distributed data, t indicates multivariate t-distributed data, $\bar{c}^2$ = average estimate of c^2 after 1000 simulations, $\bar{SE}$ =the average of the c^2-associated standard errors for all 1000 simulations, SE the true standard error of all the 1000 estimates of c^2, and Cov = coverage of the true parameter σ_c^2 within the 1000 95% constructed confidence intervals across the datasets.)

that the distribution of the traits are not normal (by definitio they are BLGP in this instance) as the models (especially the NACE) would expect teh traits to be.

While Table 3 illustrated that both models had the same relationship, Table 4 illustrates the same finding as in Table 3 for the multivariate t-distributed (as the samples increase, the standard error average and the true standard error decrease); however, the coverage rate for the ACE model is roughly stagnant, whereas the coverage rate for Falconer's model actually decreases with the sample size. This finding may have to do with how the twins are sampled from, and the fact that the ordering of the twin pairs may matter, but both models assume that it does not. This is a hard to understand finding.

Finally, for Table 5, it was found that as the estimates for both models were already essentially converged, and that the standard error average and the 'true' standard error decrease and converge to one another as the sample size increases. Furthermore, the coverage rate increases almost to the max value for the ACE model as the sample size increases, whereas the coverage rate for Falconer's model stays stagnant (but already at a very high coverage rate). This makes sense fundamentally for the ACE model, given the essential assumption of the normality of the trait being satisfied.

The results found herein in this simulation (and additional simulations compared to the paper [1] which is being validated) demonstrated the same trends (and almost exactly the same parameters) for many of the NACE model output. Outside of that, for Falconer's, the estimates, average SE and 'true' SE relationships also validated the paper. Even the coverage rate followed the same relationship of the paper–with the exception that, for an unexplained reason, the coverage rates were many times greater than the ACE model which contradicts the paper. This is an open question and is worthy of an investigation. The only explanation that may be immediately obvious is a different implementation of Falconer's formula than would be anticipated.

6. CONCLUSION

There were a few main takeaways from the project. 1) simulation is a powerful tool for ascertaining the performance of different models on distributions of interest, and especially for seeing when models fail due to violations in the model assumptions basd on the simulated data. Essentially, simulation can determine the edge-cases of when a model may work well or not. 2) As the sample size increases in these sets of models, the average standard error and true error decrease and in the ideal situation (when the data is normally distributed) converge if they have not already done so. 3) as the distribution changes from non-normal to normal traits, both models perform better in terms of the coverage rate (increases from BLGP to t to normally distributed data to almost the maximum value) and in terms of the average standard error and the 'true' standard error (they converge to the same value the more normal the distribution becomes, and become lower). 4) When the model assumptions for ACE are violated (non-normality, unequal variances across MZ and DZ groups) the model results in more severely biased estimates or coverage. 5) the coverage rate for Falconer's model appears to contradict that found in [1], otherwise the results strongly converge to that of [1]. 6) The delta method is a suitable means for creating a confidence interval when the goal is to transform the data (in this case the variance component to the heritability estimate, etc.), as it transforms the initial standard error to the

BLGP-Dist	ACE h^2	Falconer h^2
MZ/DZ Pairs	$(\bar{h}^2, \bar{SE}, SE, Cov.)$	$(\bar{h}^2, \bar{SE}, SE, Cov.)$
N = 50	(0.48,0.14,0.29,0.61)	(0.51,0.22,0.39,0.73)
N = 100	(0.50,0.12,0.24,0.64)	(0.51,0.16,0.29,0.70)
N = 200	(0.49,0.09,0.20,0.63)	(0.50,0.11,0.22,0.67)
N = 400	(0.50,0.07,0.14,0.63)	(0.50,0.08,0.15,0.70)
N = 700	(0.50,0.05,0.12,0.63)	(0.51,0.06,0.12,0.67)

Table 3: Varying twin pair sample size across BLGP distributed data in 1000 simulations across MZ/DZ pairs = 50, 100, 200, 400, and 700.
(ACE = NACE, Falc = Falconer's Formula, blgp = bivariate lagrangian poisson distributed data, t indicates multivariate t-distributed data, $\bar{h}^2$ = average estimate of h^2 after 1000 simulations, $\bar{SE}$ =the average of the h^2-associated standard errors for all 1000 simulations, SE the true standard error of all the 1000 estimates of h^2, and Cov = coverage of the true parameter σ_a^2 within the 1000 95% constructed confidence intervals across the datasets.)

new space. 7) The Anderson-Darling test is a suitable means to test for normality, while upon searching the literature, there appears to be only very recent developments in looking for equivalent tests for multivariate t-distributed data (ignoring the BLGP distribution). Overall, simulation allowed investigation into all these factors, so its utility as a tool for future research cannot be discounted.

Future work should focus on using simulation to evaluate models based on their assumptions and differing distributions as in this instance. Additionally, it should be determined why there was a notable difference in Falconer's formula coverage results as opposed to that found in [1], given all other results were consistent. It may be interesting to look into other multivariate distributions as well, in the context of these models, and see how other data are typically distributed to see how well that trait data would fare when analyzing the data with these models through future simulation.

7. REFERENCES

[1] Arbet, J., McGue, M., & Basu, S. (2020). A robust and unified framework for estimating heritability in twin studies using generalized estimating equations. Statistics in Medicine.
[2] Hofert, M. (2013). On sampling from the multivariate t distribution. The R Journal, 5(2), 129-136.
[3] Fleiss, J. L., Levin, B., & Paik, M. C. (2013). Statistical methods for rates and proportions. John Wiley & Sons.
[4] Keener, R. W. (2011). Theoretical statistics: Topics for a core course. Springer.

t-Dist	ACE h^2	Falconer h^2
MZ/DZ Pairs	$(\bar{h}^2, \bar{SE}, SE, Cov.)$	$(\bar{h}^2, \bar{SE}, SE, Cov.)$
N = 50	(0.47,0.16,0.24,0.72)	(0.51,0.22,0.34,0.82)
N = 100	(0.48,0.12,0.20,0.76)	(0.49,0.16,0.24,0.80)
N = 200	(0.50,0.09,0.16,0.74)	(0.50,0.11,0.18,0.78)
N = 400	(0.50,0.07,0.12,0.73)	(0.50,0.08,0.13,0.79)
N = 700	(0.50,0.05,0.10,0.73)	(0.50,0.06,0.11,0.74)

Table 4: Varying twin pair sample size across multivariate t-distributed data in 1000 simulations across MZ/DZ pairs = 50, 100, 200, 400, and 700.
(ACE = NACE, Falc = Falconer's Formula, blgp = bivariate lagrangian poisson distributed data, t indicates multivariate t-distributed data, $\bar{h}^2$ = average estimate of h^2 after 1000 simulations, $\bar{SE}$ =the average of the h^2-associated standard errors for all 1000 simulations, SE the true standard error of all the 1000 estimates of h^2, and Cov = coverage of the true parameter σ_a^2 within the 1000 95% constructed confidence intervals across the datasets.)

Normal-Dist	ACE h^2	Falconer h^2
MZ/DZ Pairs	$(\bar{h}^2, \bar{SE}, SE, Cov.)$	$(\bar{h}^2, \bar{SE}, SE, Cov.)$
N = 50	(0.50,0.18,0.18,0.88)	(0.51,0.22,0.23,0.95)
N = 100	(0.50,0.13,0.13,0.95)	(0.50,0.16,0.15,0.95)
N = 200	(0.50,0.10,0.10,0.94)	(0.50,0.11,0.11,0.94)
N = 400	(0.50,0.07,0.07,0.96)	(0.50,0.08,0.08,0.95)
N = 700	(0.50,0.05,0.05,0.96)	(0.50,0.06,0.06,0.95)

Table 5: Varying twin pair sample size across bivariate normal distributed data in 1000 simulations across MZ/DZ pairs = 50, 100, 200, 400, and 700.
(ACE = NACE, Falc = Falconer's Formula, blgp = bivariate lagrangian poisson distributed data, t indicates multivariate t-distributed data, $\bar{h}^2$ = average estimate of h^2 after 1000 simulations, $\bar{SE}$ =the average of the h^2-associated standard errors for all 1000 simulations, SE the true standard error of all the 1000 estimates of h^2, and Cov = coverage of the true parameter σ_a^2 within the 1000 95% constructed confidence intervals across the datasets.)