1 SUPPLEMENTARY MATERIALS

1.1 PROOF OF PROPOSITION 3.4

Proposition 3.4 $\mathbf{x} \sim \mathcal{N}_d(\mu, \mathbf{\Sigma})$ *if and only if* $\mathbf{z} = \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \mu) \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$.

Proof. By eigendecomposition, $\mathbf{\Sigma}^{-1} = U\Lambda U^T$. Define $\mathbf{\Sigma}^{-\frac{1}{2}} = U\Lambda^{\frac{1}{2}}U^T$. Then $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}^{-\frac{1}{2}}$.

$\Leftarrow$
Let $\mathbf{z} = \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \mu) \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$, then $\mathbf{x} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{z} + \mu$. Since $\mathbf{x}$ is a linear transformation of $\mathbf{z}$ then $\mathbf{x}$ follows multivariate normal distribution with dimension d.

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \mathbf{\Sigma}^{\frac{1}{2}}\mathbb{E}[\mathbf{z}] + \mu = \mu \\ Cov(\mathbf{x}) &= Cov(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{z} + \mu) = Cov(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{z}) \\ &= \mathbf{\Sigma}^{\frac{1}{2}}Cov(\mathbf{z})\mathbf{\Sigma}^{\frac{1}{2}^T} = \mathbf{\Sigma}\end{aligned}$$

Therefore, $\mathbf{x} \sim \mathcal{N}_d(\mu, \mathbf{\Sigma})$.
$\Rightarrow$
Since $\mathbf{z}$ is a linear transformation of $\mathbf{x}$, then $\mathbf{z}$ follows multivariate normal with dimension d.

$$\begin{aligned}\mathbb{E}[\mathbf{z}] &= \mathbf{\Sigma}^{-\frac{1}{2}}(\mathbb{E}[\mathbf{x}] - \mu) = \mathbf{0} \\ Cov(\mathbf{z}) &= Cov(\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \mu)) = Cov(\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x}) \\ &= \mathbf{\Sigma}^{-\frac{1}{2}}Cov(\mathbf{x})\mathbf{\Sigma}^{-\frac{1}{2}^T} \\ &= U\Lambda^{\frac{1}{2}}U^T(U\Lambda U^T)U\Lambda^{\frac{1}{2}}U^T \\ &= \mathbf{I}.\end{aligned}$$

Therefore, $\mathbf{z} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$. □

1.2 GRADIENT OF W

Proposition 4.1 *Let k be the size of the mini-batch. For any layer ℓ, denote $\mathbf{Y}^\ell = \mathbf{\Theta}^\ell\mathbf{Y}^{(\ell-1)}$, where $\mathbf{Y}^{(\ell-1)}$ is an $(n \times k)$ matrix of arbitrary activation from layer $(\ell - 1)$, $\mathbf{Y}^\ell$ is an $(m \times k)$ matrix of linear activation for layer ℓ, and $\mathbf{\Theta}^\ell$ is the $(m \times n)$ matrix of parameters connecting the layers. Let $\mathbf{y}$ be the $(mk \times 1)$ ascendingly sorted vectorization of $\mathbf{Y}^\ell$. Then, $\mathbf{y}$ can be computed by $\mathbf{A}\theta$ where $\mathbf{A}$ and θ are the re-organized $\mathbf{Y}^{\ell-1}$ and the vectorization of $\mathbf{\Theta}^\ell$. Specifically, $\mathbf{A}$ is an $(mk \times mn)$ matrix with each row containing the relevant $\mathbf{Y}^{(\ell-1)}$ data for a particular node's activation. The gradient of W can be computed by*

$$\nabla_\theta W = \frac{2\mathbf{a}^T\mathbf{A}\theta}{\theta^T\mathbf{Z}\theta}\mathbf{a}^T\mathbf{A}\left[\mathbf{I} - \frac{\theta\theta^T\mathbf{Z}}{\theta^T\mathbf{Z}\theta}\right], \qquad (6)$$

where $\mathbf{Z} = \mathbf{A}(\mathbf{I} - \frac{\mathbf{J}}{mk})\mathbf{A}^T$, $\mathbf{I}$ is mk-dimensional identity matrix, and $\mathbf{J}$ is a $(mk \times mk)$ matrix of ones.

Proof. Following the notation, then

$$W = \frac{(\mathbf{a}^T\mathbf{y})^2}{||(\mathbf{I} - \frac{\mathbf{J}}{mk})\mathbf{y}||_2^2}.$$

Letting $\mathbf{Z} = \mathbf{A}^T(\mathbf{I} - \frac{\mathbf{J}}{mk})\mathbf{A}$, we have

$$W = \frac{(\mathbf{a}^T\mathbf{A}\theta)^2}{\theta^T\mathbf{Z}\theta}$$

$$\nabla_\theta W = \frac{2\mathbf{a}^T\mathbf{A}\theta\mathbf{a}^T\mathbf{A}}{\theta^T\mathbf{Z}\theta} - \frac{2\mathbf{a}^T\mathbf{A}\theta\mathbf{a}^T\mathbf{A}\mathbf{a}^T\mathbf{A}\theta\theta^T\mathbf{Z}}{(\theta^T\mathbf{Z}\theta)^2}$$

Simplifying the above formula yields Eq.(6). □

1.3 CONDITION VERIFICATION FOR BOTTOU'S THEOREM

The test statistic W used for the loss, is differentiable almost everywhere. Specifically, the first three derivatives exist, and furthermore the second and third derivative are continuous almost everywhere. By equivalent form of the the regularizer $\lambda||\theta||^2$, then $||\theta||^2 \leq \eta$ where η is some proper value corresponding to λ. Therefore, boundedness for the second and third derivatives follows from the fact that a continuous function is bounded on a fixed region. The four assertions are verified below.

(i) Choose $H(w) = \nabla_\theta W$

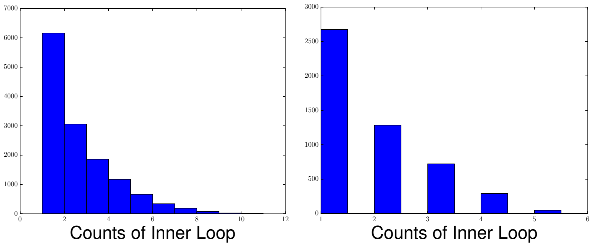
(ii) Typical assumption on the learning rate, satisfied by Adam (Kingma and Ba, 2014).

(iii) Let $\mathbf{q} = \frac{\mathbf{a}^T\mathbf{A}\theta\mathbf{a}^T\mathbf{A}}{\theta^T\mathbf{Z}\theta}$, thus $W = \mathbf{q}\theta$. Then
$H(\theta) = \nabla_\theta W = 2\mathbf{q}\Big[I - \frac{\theta\theta^T\mathbf{Z}}{\theta^T\mathbf{Z}\theta}\Big] = \Big(2\mathbf{q} - \frac{2W\theta^T\mathbf{Z}}{\theta^T\mathbf{Z}\theta}\Big)$
Then
$\mathbb{E}(H(\theta)^2) = \Big(2\mathbf{q} - \frac{2W\theta^T\mathbf{Z}}{\theta^T\mathbf{Z}\theta}\Big)\Big(2\mathbf{q} - \frac{2W\theta^T\mathbf{Z}}{\theta^T\mathbf{Z}\theta}\Big)^T$
$= (2\mathbf{q} - \frac{2W}{\theta^T\mathbf{Z}\theta}\theta^T\mathbf{Z})(2\mathbf{q}^T - \frac{2W}{\theta^T\mathbf{Z}\theta}\mathbf{Z}^T\theta)$
$= 4\mathbf{q}\mathbf{q}^T - \frac{4W}{\theta^T\mathbf{Z}\theta}(\mathbf{q}\mathbf{Z}^T\theta + \theta^T\mathbf{Z}\mathbf{q}^T) + \frac{4W^2}{(\theta^T\mathbf{Z}\theta)^2}\theta^T\mathbf{Z}\mathbf{Z}^T\theta$
$\leq 4\mathbf{q}\mathbf{q}^T - \frac{4W}{\theta^T\mathbf{Z}\theta}(\mathbf{q}\mathbf{Z}^T\theta + \theta^T\mathbf{Z}\mathbf{q}^T) + \frac{4W}{(\theta^T\mathbf{Z}\theta)^2}\theta^T\mathbf{Z}\mathbf{Z}^T\theta$
Take $A = 4\mathbf{q}\mathbf{q}^T$, $B = \Big(\frac{4}{(\theta^T\mathbf{Z}\theta)^2}\theta^T\mathbf{Z}\mathbf{Z}^T\theta - \frac{4}{\theta^T\mathbf{Z}\theta}(\mathbf{q}\mathbf{Z}^T\theta + \theta^T\mathbf{Z}\mathbf{q}^T)\Big)$, and $C(w) = W$.

(iv) Shapiro-Wilk W is bounded between 0 and 1, and using lemma 3.3, from (Shapiro and Wilk, 1965), the minimum value of W is $\frac{na_1^2}{(n-1)}$.

1.4 INNER LOOP ITERATIONS FOR HTAE-(M)SW

As can be seen in Fig.3, with 10^5 reconstruction iterations, the encoder was able to induce failure to reject normality in a small number of iterations. This seemed to hold as well for other tests shown in the following section.



(a) Univariate

(b) Multivariate

Figure 3: Number of iterations needed to bring $q(\mathbf{z})$ back into class $\mathcal{G}$ in both univariate and multivariate hypothesis tests.

1.5 OTHER GOODNESS-OF-FIT TESTS

After conducting a simulation study based on type I and type II error (Mecklin and Mundfrom, 2005) concluded that no single test was best in all situations. However, the authors did suggest using Royston's (Royston, 1983) and Henze-Zirkler's tests (Henze and Zirkler, 1990) because of good power and type I error control. These tests as well as the Malkovich-Afifi test (Malkovich and Afifi, 1973), and Mardia's Skewness test (Mardia, 1970) were reviewed and substituted into Alg. 1 with appropriate changes to the sign of the loss. Several of these methods for assessing normality can be seen in (Korkmaz et al., 2014), and are reviewed immediately, along with other techniques not presented here. As described in the experiments each of the following tests used the same α, network architecture, Dropout parameters, and batch size. Generated images for these methods along with their inner loop empirical count distributions can be seen in Fig. 4. Run times for our approach with these additional hypothesis tests are included in Table 2.

1.5.1 Royston's H Test

To test for multivariate normality, Royston's test uses either the Shapiro-Wilk (Shapiro and Wilk, 1965) or the Shapiro-Francia (Shapiro and Francia, 1972) statistic. When the kurtosis of the data is greater than 3, the Shapiro-Francia test is used. Otherwise, the Shapiro-Wilk test is used.

$$w_j = \begin{cases} W_j^\dagger & \text{Kurt}[X] > 3 \\ W_j & \text{otherwise,} \end{cases}$$

where W_j is the Shapiro-Wilk statistic and $W_j^\dagger$ is the Shapiro-Francia statistic for the j^{th} variable. The number of samples dictates the next step:

$$\begin{array}{lll} 4 \leq n \leq 11 & x = n & w_j = -log[\gamma - log(1 - W_j)] \\ 12 \leq n \leq 2000 & x = log(n) & w_j = log(1 - W_j) \end{array}$$

Define Z_j from the normality transformation in (Royston, 1992) to be $Z_j = \frac{w_j - \mu}{\sigma}$. γ, μ, σ are derived from polynomial approximations where the coefficients are given in (Royston, 1992) for different samples sizes.

$$\gamma = a_{0_\gamma} + a_{1_\gamma} x + a_{2_\gamma} x^2 + \ddots + a_{d_\gamma} x^d$$

$$\mu = a_{0_\mu} + a_{1_\mu} x + a_{2_\mu} x^2 + \ddots + a_{d_\mu} x^d$$

$$log(\sigma) = a_{0_\sigma} + a_{1_\sigma} x + a_{2_\sigma} x^2 + \ddots + a_{d_\sigma} x^d$$

Royston's test statistic, H, for multivariate normality is defined as

$$H = \frac{e \sum_{j=1}^{p} \psi_j}{p} \sim \chi_e^2,$$

where e is the degrees of freedom defined to be $e = \frac{p}{[1 + (p-1)\bar{c}]}$ and ψ_j is defined to be $\psi_j = (\Phi^{-1}[\Phi(\frac{-Z_j}{2})])^2 \quad j = 1, 2, ..., p$. $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution, and $\Phi(\cdot)^{-1}$ its inverse. Let R be the correlation matrix, and r_{ij} be the correlation between the i^{th} and j^{th} variables, then $\bar{c}$ is defined as $\bar{c} = \sum_j \sum_j \frac{c_{ij}}{p(p-1)} \quad i \neq j$ where

$$c_{ij} = \begin{cases} g(r_{ij}, n), & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases}$$

with the boundaries of $g(\cdot)$ as $g(0, n) = 0$ and $g(1, n) = 1$. The function $g(\cdot)$ is defined as $g(r, n) = r^\lambda[1 - \frac{\mu}{\nu}(1 - r)^\mu]$. The unknown parameters μ, λ, and ν were estimated from a simulation (Ross et al., 1980) where $\mu = 0.715$ and $\lambda = 5$ for sample size $10 \leq n \leq 2000$. The parameter ν is a cubic function which is obtained by $\nu(n) = 0.21364 + 0.015124x^2 - 0.0018034x^3$ where $x = log(n)$. From the hypothesis test if $H > H_\alpha$ where H_α is the critical value, then the null hypothesis is rejected, in favor of the alternative, that the data *not* multivariate normal. The HTAE using Royston's H test as the critic will be denoted HTAE-R.

1.5.2 Malkovich-Afifi Test

Another approach to testing for multivariate normality was presented by (Malkovich and Afifi, 1973) where the authors made use of Roy's union-intersection principle (Roy, 1953). The union-intersection approach to hypothesis testing can be used to express the null as an intersection. For example, it is possible to denote

$H_0 : \theta \in \cap_{\omega \in \Omega} \Theta_\omega$ where Ω is an index set that may or may not be finite. If for every $\omega \in \Omega$ the null is not rejected, then $H_{0\omega}$ is not rejected. However, if a single $H_{0\omega}$ is rejected, the null is rejected. By following Roy's union-intersection principle of test construction, H_0 will be accepted when univariate normality of the projected samples $c'X$ is accepted for any $c \neq 0$. Thus, for a particular c we can let $Y_j = c'X_j$ for $j = 1, ..., n$ and let $Y_{(1)}, ..., Y_{(n)}$ denote the order statistics of the Y_j's. The Shapiro-Wilk form, dependent on c, can be written as $W(c) = \frac{(\sum_{i=1}^n a_i (Y_{(i)} - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$, where $\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$ and the a_i's are the same from (Shapiro and Wilk, 1965). By using Roy's union intersection principle (Roy, 1953), (Malkovich and Afifi, 1973) combined $W(c), c \in \mathcal{C}$ to define the test $W_{MA} = \inf_{\forall c \in \mathcal{C}} W(c)$. As in the Shapiro-Wilk statistic, a small value signifies rejection of multivariate normality. Stated another way, the MA test of multivariate normality under H_0 fails to reject the null hypothesis if $\min_c W(c^T X_1, c^T X_2, ..., c^T X_n) \geq K_w$ where K_w is a constant. Direct numerical evaluation is not possible so the authors proposed another approach by noting that $W(c^T X_1, c^T X_2, ..., c^T X_n)$ has a lower bound when c^T satisfies the conditions of (Shapiro and Wilk, 1965): $c^T(X_l - \bar{X}) = \frac{n-1}{na_1}, \quad c^T(X_j - \bar{X}) = -\frac{1}{na_1}$for $j = 1, ..., n$ and $j \neq l$ where $\bar{X}$ is the mean vector. A solution c for these equations does not exist, so the authors instead find a vector c^T which minimizes $\left[c^T(X_l - \bar{X}) - \frac{n-1}{na_1}\right]^2 + \sum_{j \neq l}^n \left[c^T(X_j - \bar{X}) + \frac{1}{na-1}\right]^2$. The vector $c^{(l)} = \frac{1}{a_1} A^{-1}(X_l - \bar{X})$ and $A = \sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})^T$. As $c^{(l)}$ may be any of the $l \in \{1, 2, ..., n\}$ vectors, the authors chose $c^{(m)} \in \{c^{(1)}, c^{(2)}, ..., c^{(n)}\}$ such that the denominator of $W(c^T X_1, c^T X_2, ..., c^T X_n)$ is maximized over these n choices. Formally, $(X_m - \bar{X})^T A^{-1}(X_m - \bar{X}) = \max_{1 \leq l \leq n}(X_l - \bar{X})^T A^{-1}(X_l - \bar{X})$. Pseudo-code for computing the test statistic W_{MA} is detailed in Alg.2. (Fattorini, 1986) proposed a modification of the Malkovich-Afifi statistic, into $FA(X_1, X_2, ..., X_n) = \min_{1 \leq l \leq n} W(c^{(l)^T} X_1, c^{(l)^T} X_2, ..., c^{(l)^T} X_n)$. The Malkovich-Afifi (MA) and Fattorini (FA) hypothesis tests reject the null hypothesis of multivariate normality for small values of the statistic. Unfortunately, the null distribution of W_{MA} is not known, and so simulation is necessary to identify empirical critical values prior to using Alg.1. For our experiments 10,000 batches of size 100 were sampled from $\mathcal{N}_8(0, \mathbf{I})$. The empirical significance point associated with $\alpha = 0.05$ was used as the critical value for the inner loop, i.e. the inner loop became active when the test statistic was less than this value. The HTAE using the Malkovich-Afifi test as the critic will be denoted HTAE-MA.

Algorithm 2 Malkovich-Afifi MVN UIT Test Statistic

```
Input: $X_1, ..., X_n$
$X_m = \arg\max_{1 \leq l \leq n}(X_l - \bar{X})^T A^{-1}(X_l - \bar{X})$
Calculate $U_j = (X_m - \bar{X})^T A^{-1}(X_j - \bar{X}) \quad j = 1, 2, ..., n$
Denote sorted statistics $U_j$ as $U_{(1)}, U_{(2)}, ..., U_{(m)}$
$W_{MA} = \frac{\left[\sum_{j=1}^n a_j U_{(j)}\right]^2}{(X_m - \bar{X})^T A^{-1}(X_m - \bar{X})}$ {where $a_j$ are constants from (Shapiro and Wilk, 1965)}
```

1.5.3 Mardia's Skewness Test

In (Mardia, 1970) a multivariate test for normality is proposed based on multivariate extensions of skewness and kurtosis. The skewness test statistic is defined as

$$\hat{\gamma}_{1,p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n m_{ij}^3,$$

where $m_{ij} = (x_i - \bar{x})^T S^{-1}(x_j - \bar{x})$, $S = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(x_k - \bar{x})^T$, and p are the number of features. Under the null hypothesis, that the data come from a multivariate normal distribution, $\frac{n}{6}\hat{\gamma}_{1,p} \sim \chi^2_{df}$ where degrees of freedom, $df = \frac{p(p+1)(p+2)}{6}$. In (Mardia, 1974) a correction for small samples, $n < 20$, was incorporated into the test. The small sample test statistic is defined to be $\frac{nb}{b}\hat{\gamma}_{1,p}$ with $b = \frac{(p+1)(n+1)(n+3)}{(n(n+1)(p+1)-6)}$ and is distributed as χ^2_{df} where df is defined as before. There are still problems with this test, specifically it is not consistent against symmetric non-normal alternatives (Baringhaus and Henze, 1991). The HTAE using Mardia's skewness test as the critic will be denoted HTAE-M.

1.5.4 Henze-Zirkler's Test

(Henze and Zirkler, 1990) proposed a multiviariate normality test based on a non-negative functional distance. The test statistic is defined as

$$HZ = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n e^{-\frac{\beta^2}{2} D_{ij}}$$

$$-2(1 + \beta^2)^{-\frac{p}{2}} \sum_{i=1}^n e^{-\frac{\beta^2}{2(1+\beta^2)} D_i} + n(1 + 2\beta^2)^{-\frac{p}{2}}$$

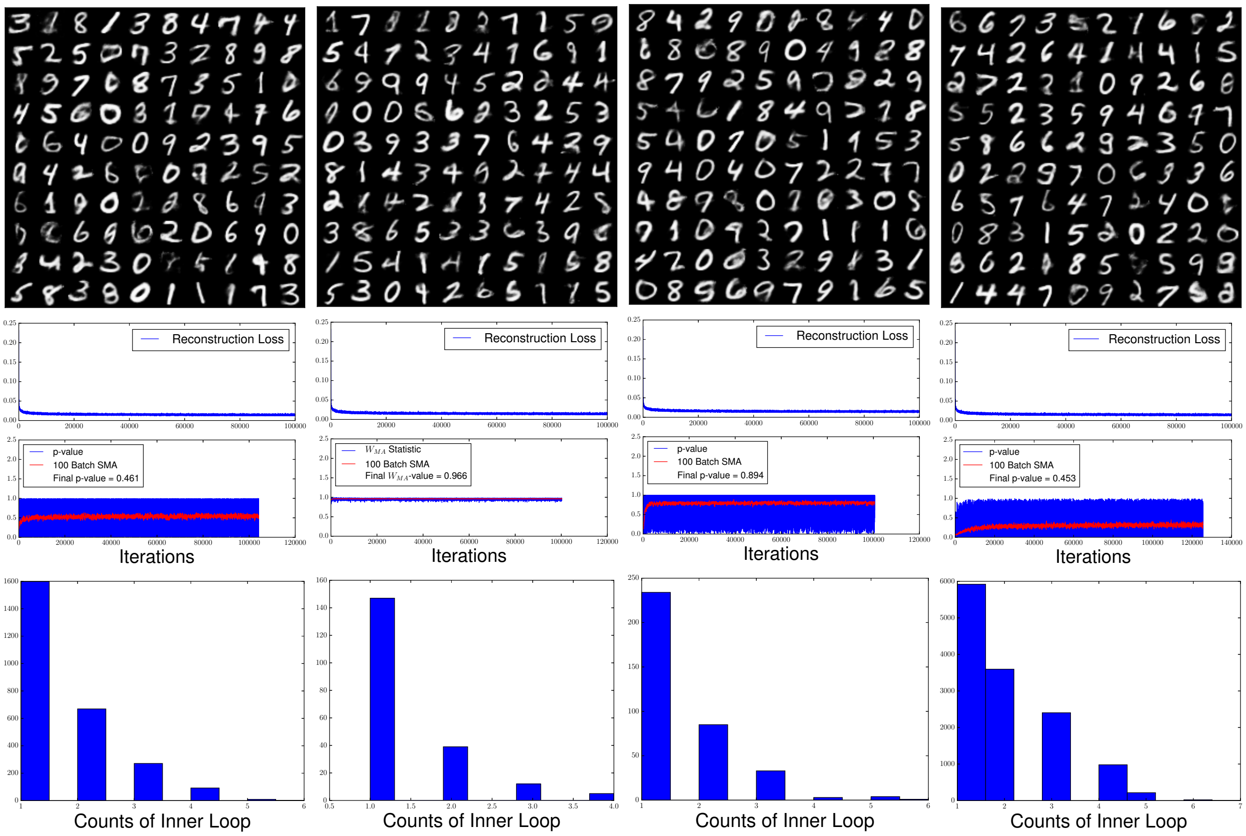
where $D_i - (x_i - \bar{x}^T S^{-1}(x_i - \bar{x})$, $D_{ij} = (x_i - x_j)^T S^{-1}(x_i - x_j)$, $\beta = \frac{1}{sqrt2}\left(\frac{n(2p+1)}{4}\right)^{(\frac{1}{p+4})}$, and p is the number of features. Under the null hypothesis the test statistic HZ is approximately log-normally distributed with mean μ and variance σ^2 defined as such

$$\begin{aligned}\mu &= 1 - \frac{a^{-\frac{p}{2}}\left(1 + p\beta^{\frac{2}{a}} + (p(p+2)\beta^4)\right)}{2a^2} \\ \sigma^2 &= 2(1+4\beta^2)^{-\frac{p}{2}} + \frac{2a^{-p}(1+2p\beta^4)}{a^2} + \frac{3p(p+2)\beta^8}{4a^4} \\ &\quad - 4w_\beta^{-\frac{p}{2}}\left(1 + \frac{3p\beta^4}{2w_\beta} + \frac{p(p+2)\beta^8}{2w_\beta^2}\right),\end{aligned}$$

where $a = 1{+}2\beta^2$ and $w_\beta = (1{+}\beta^2)(1{+}3\beta^2)$. Defining $\mu_{log} = log\left(\sqrt{\frac{\mu^4}{\sigma^2+\mu^2}}\right)$ and $\sigma_{log}^2 = log\left(\frac{\sigma^2+\mu^2}{\mu^2}\right)$. The final Wald test statistic is given as $\frac{log(HZ)-\mu_{log}}{\sigma_{log}}$. The HTAE using the Henze-Zirkler test as the critic will be denoted HTAE-HZ.

Table 2: Run time in seconds for 10^5 iterations for the 8-dimension multivariate cases using an NVIDIA GTX 1080Ti GPU.

Method	8-D
HTAE-R	1429.77
HTAE-M	550.61
HTAE-MA	1356.86
HTAE-HZ	734.84



(a) HTAE-R (b) HTAE-MA (c) HTAE-M (d) HTAE-HZ

Figure 4: The top row plots random samples generated from each HTAE model. Row two contains the p-values with a 100 batch SMA for HTAE-R, HTAE-M, and HTAE-HZ. The HTAE-MA model used the critical value in place of the p-value. Row three are frequency counts of inner loops necessary to ensure failure to reject was met.

References

L. Baringhaus and N. Henze. Limit distributions for measures of multivariate skewness and kurtosis based on projections. *Journal of Multivariate Analysis*, 38(1): 51–69, 1991.

L. Fattorini. Remarks on the use of shapiro-wilk statistic for testing multivariate normality. *Statistica*, 46(2):209–217, 1986.

N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10):3595–3617, 1990.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Korkmaz, D. Goksuluk, and G. Zararsiz. Mvn: An r package for assessing multivariate normality. *The R Journal*, 6(2):151–162, 2014.

J. F. Malkovich and A. Afifi. On tests for multivariate normality. *Journal of the american statistical association*, 68(341):176–179, 1973.

K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

K. V. Mardia. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 115–128, 1974.

C. J. Mecklin and D. J. Mundfrom. A monte carlo comparison of the type i and type ii error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, 75(2):93–107, 2005.

G. Ross, R. Jones, R. Kempton, F. Laukner, R. Payne, D. Hawkins, and R. White. *MLP: maximum likelihood program*. 1980.

S. N. Roy. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, pages 220–238, 1953.

J. Royston. Some techniques for assessing multivarate normality based on the shapiro-wilk w. *Applied Statistics*, pages 121–133, 1983.

P. Royston. Approximating the shapiro-wilk w-test for non-normality. *Statistics and Computing*, 2(3):117–119, 1992.

S. S. Shapiro and R. Francia. An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337):215–216, 1972.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4): 591–611, 1965.