The data is paired, clustered, and balanced: X_{ij} consisting of $j = 1, \ldots, m_i$, real values on each of $i = 1, \ldots, N$ clusters, and Y_{ij} consisting of $j = 1, \ldots, n_i$ real values also on $i = 1, \ldots, N$ clusters. E.g., X_{ij} and Y_{ij} may correspond respectively to non-diseased and diseased measurement j on subject i. One statistic for measuring the AUC given by the measurements X_{ij}, Y_{ij} averages the Mann-Whitney non-parametric AUC estimate of the clusters:

$$U_N = \frac{1}{N}\sum_{i=1}^{N} H_{ii} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{n_i m_i}\sum_{j,k}\{X_{ij} < Y_{ik}\},. \tag{1}$$

Here $H_{ij} := \frac{1}{n_i m_j}\sum_{j,k}\{X_{ij} < Y_{ik}\}$ is the AUC statistic computed using the non-diseased measurements of cluster i and the diseased measurements of cluster j. Another statistic computes the non-parametric estimate ignoring the cluster structure,

$$V_N = \frac{1}{N^2}\sum_{i,j} H_{ij} = \frac{1}{\sum m_i \sum n_i}\sum_{i,j}\sum_{r,s}\{X_{ir} < Y_{js}\}. \tag{2}$$

These are almost but not quite U-statistics because of the dependence among observations (diseased and non-diseased) within a cluster. To compute the asymptotic behavior I made the following assumptions about the data.

1. Clusters are i.i.d., i.e., $dF_{X_{ir},Y_{js}} = dF_{X_{ir}} dF_{Y_{js}}$ and $dF_{X_{ir},X_{js}} = dF_{X_{ir}} dF_{X_{js}}$ when $i \neq j$[[pairwise or joint?]]
2. Within a cluster, the X_{ij} values are exchangeable, as are the Y_{ij} values within a cluster
3. Within a cluster, the joint distribution of X_{ij} and Y_{ij} values is fixed and constant across clusters

Asymptotic distribution of U_N. U_N is a sum of i.i.d. variables of bounded variance. Let $\theta_{11} = \mathbb{E}[U_N] = \mathbb{E}[H_1 1] = \mathbb{P}[X_{11} < Y_{11}]$ denote the probability that a non-diseased value is less than a diseased value in the same cluster. By the CLT,

$$\frac{\sum_{i=1}^{N} H_{ii} - N\theta_{11}}{\sqrt{VarH_{11}}} \rightsquigarrow N(0,1),$$

and so,

$$\sqrt{N}(U_N - \theta_{11}) \rightsquigarrow N(0, VarH_{11}).$$

The asymptotic variance above is computed as

$$\begin{aligned} VarH_{11} =& \frac{1}{mn}(\mathbb{P}[X_{11} < Y_{11}] + (n-1)\mathbb{P}[X_{11} < Y_{11}, X_{11} < Y_{12}] + (m-1)\mathbb{P}[X_{11} < Y_{11}, X_{12} < Y_{11}] \\ &+ (n-1)(m-1)\mathbb{P}[X_{11} < Y_{11}, X_{12} < Y_{12}]) - \mathbb{P}^2[X_{11} < Y_{11}]. \end{aligned}$$

Figure 1 displays the results of a simulation on data with uniform marginals.

Asymptotic distribution of V_N. Assume that the cluster sizes are constant and also equal for non-diseased and diseased observations, $n_i = m_i =: n$. Analogous to θ_{11}, let $\theta_{12} = \mathbb{E}[H_{12}] = \mathbb{P}[X_{11} <$

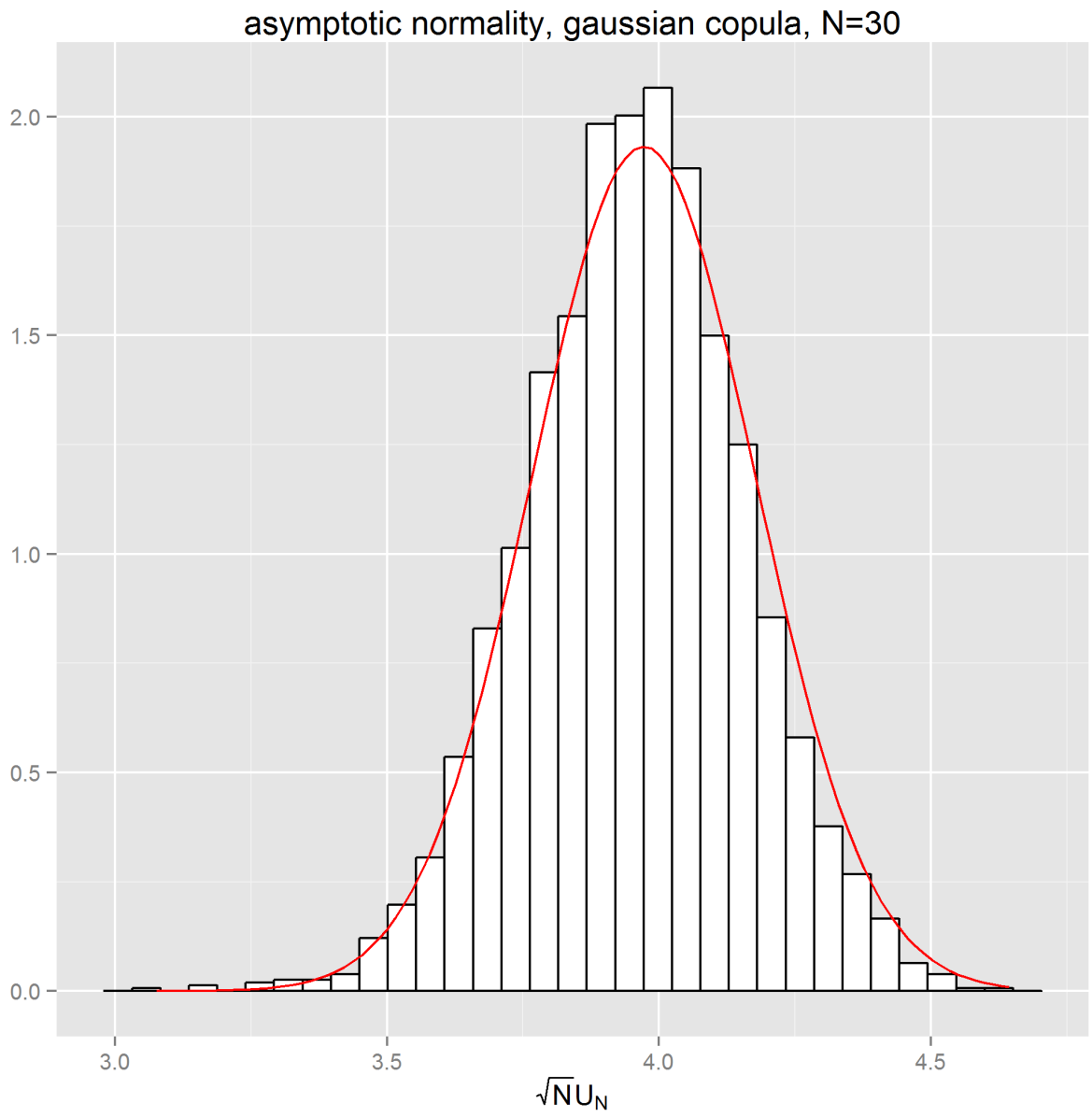

Figure 1: Histogram of 1e3 samples of $\sqrt{N}U_N$ overlaid with a $N(\sqrt{N}\widehat{\theta_{11}}, \widehat{Var(H_{11})})$ density. The hats are meant to indicate that the quantities are empirical plug-in estimates of the asymptotic mean and variance formulas given above. The data in each replicate consists of $N = 30$ pairs (X_i, Y_i) of correlated vectors of length $n = m = 20$ with uniform marginals, obtained by applying the normal CDF to correlated normal vectors.

$Y_{21}]$ denote the probability that a non-diseased value is less than a diseased value in a different cluster. Then,

$$\sqrt{N}(V_N - \mathbb{E}[V_N]) \rightsquigarrow N(0, \mathbb{E}[\alpha(X_1, Y_1)^2] - 4\theta_{12}^2),$$

where $\alpha(X_1, Y_1) := \mathbb{E}[H(X_1, Y_2)|X_1] + \mathbb{E}[H(X_2, Y_1)|Y_1]$. In terms of the comparisons of observations,

the asymptotic variance is computed to be,

$$\mathbb{E}[\alpha(X_1, Y_1)^2] - 4\theta_{12}^2 = \tag{3}$$
$$\mathbb{P}[X_{11} < Y_{21}, X_{11} < Y_{31}] + \mathbb{P}[X_{21} < Y_{11}, X_{31} < Y_{11}] + 2\mathbb{P}[X_{11} < Y_{21}, X_{31} < Y_{11}] - 4\mathbb{P}^2[X_{11} < Y_{21}].$$

The demonstration starts by writing V_N as

$$V_N = \frac{1}{N^2}\sum_{i=1}^{N} H_{ii} + \frac{N(N-1)}{N^2}W,$$

defining the statistic $W_N = 1/(N(N-1))\sum_{i\neq j} H_{ij}$ as the average of the AUCs where the non-diseased data comes from one cluster and the diseased from another cluster. W_N is the statistic we discussed the other day. Then $\theta_{12} = \mathbb{E}[W_N]$. The asymptotic distribution of V_N is equivalent to that of W_N,

$$\begin{aligned}\sqrt{N}(V_N - \mathbb{E}[V_N]) &= N^{-3/2}\sum_{i=1}^{N} H_{ii} + \sqrt{N}(\frac{N(N-1)}{N^2}W - \frac{N(N-1)}{N^2}\theta_{12}) + N^{-1/2}\theta_{11} \\ &\sim \sqrt{N}(W_N - \theta_{12}).\end{aligned}$$

To compute the asymptotic distribution of W_N, I used the method of the "Hajék projection" $\hat{W} := \sum_{i=1}^{N}\mathbb{E}[W|(X_i, Y_i)] - (N-1)\theta_{12}$. A comparison using synthetic data of the approximation of the target statistic V_N by the Hajék projections is presented in Fig. 2. One proceeds by first showing that $\mathbb{E}[(W - \hat{W})^2] = O(1/n^2)$. Then $\sqrt{N}(W - \hat{W}) = o_P(1)$ and so the asymptotic distribution of $\sqrt{N}(W - \theta_{12})$ may be obtained using that of $\sqrt{N}(\hat{W} - \theta_{12})$. Since $\hat{W}$ is an i.i.d. sum, its asymptotic distribution can be computed using the CLT. Since the variances of the summands change as N varies, the Lindeberg-Feller (triangular array) CLT is appropriate.

Figure 2 displays the results of a simulation on data with uniform marginals.

Asymptotic distribution of (U_N, V_N). The asymptotic joint distribution can be found using the Cramér-Wold device:

$$\sqrt{N}\begin{pmatrix} U_N - \theta_{11} \\ V_N - \theta_{12} \end{pmatrix} \rightsquigarrow N(0, \begin{pmatrix} 195 \\ 280 \end{pmatrix}),$$

with $\alpha(X_1, Y_1), \theta_{12}, VarH_{11}$ expanded above. Results of a simulation are presented in fig. 3.

TODO

1. extend V_N result to case of unequal number of diseased and non-diseased observations per cluster
2. extend V_N result to case where the number of observations per cluster (diseased and non-diseased) varies across clusters

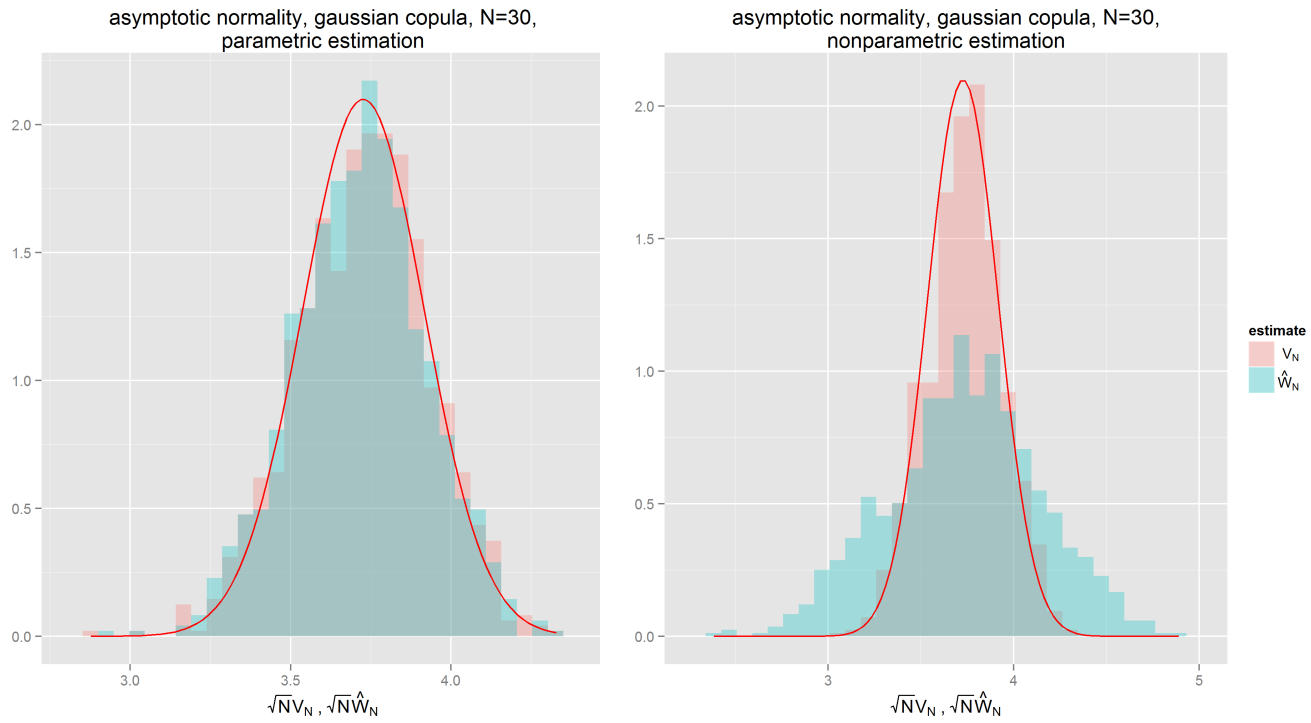

Figure 2: Histogram of 1e3 samples of $\sqrt{N}V_N$ overlaid with a $N(\sqrt{N}\widehat{\theta_{12}}, \hat{\alpha} - 4\hat{\theta_{12}}^2)$ density. The hats are meant to indicate that the quantities are empirical plug-in estimates of the asymptotic mean and variance formulas given above. The data in each replicate consists of $N = 30$ pairs (X_i, Y_i) of correlated vectors of length $n = m = 20$ with uniform marginals, obtained by applying the normal CDF to correlated normal vectors. Also shown for comparison is a histogram of 1e3 samples of $\sqrt{N}\hat{W}_N$, the i.i.d. Hajék projections used to approximate V_N.
The figure on the left is generated using a parametric procedure, in that the estimation procedure uses that the data have uniform marginals. The figure on the right uses non-parametric estimates.

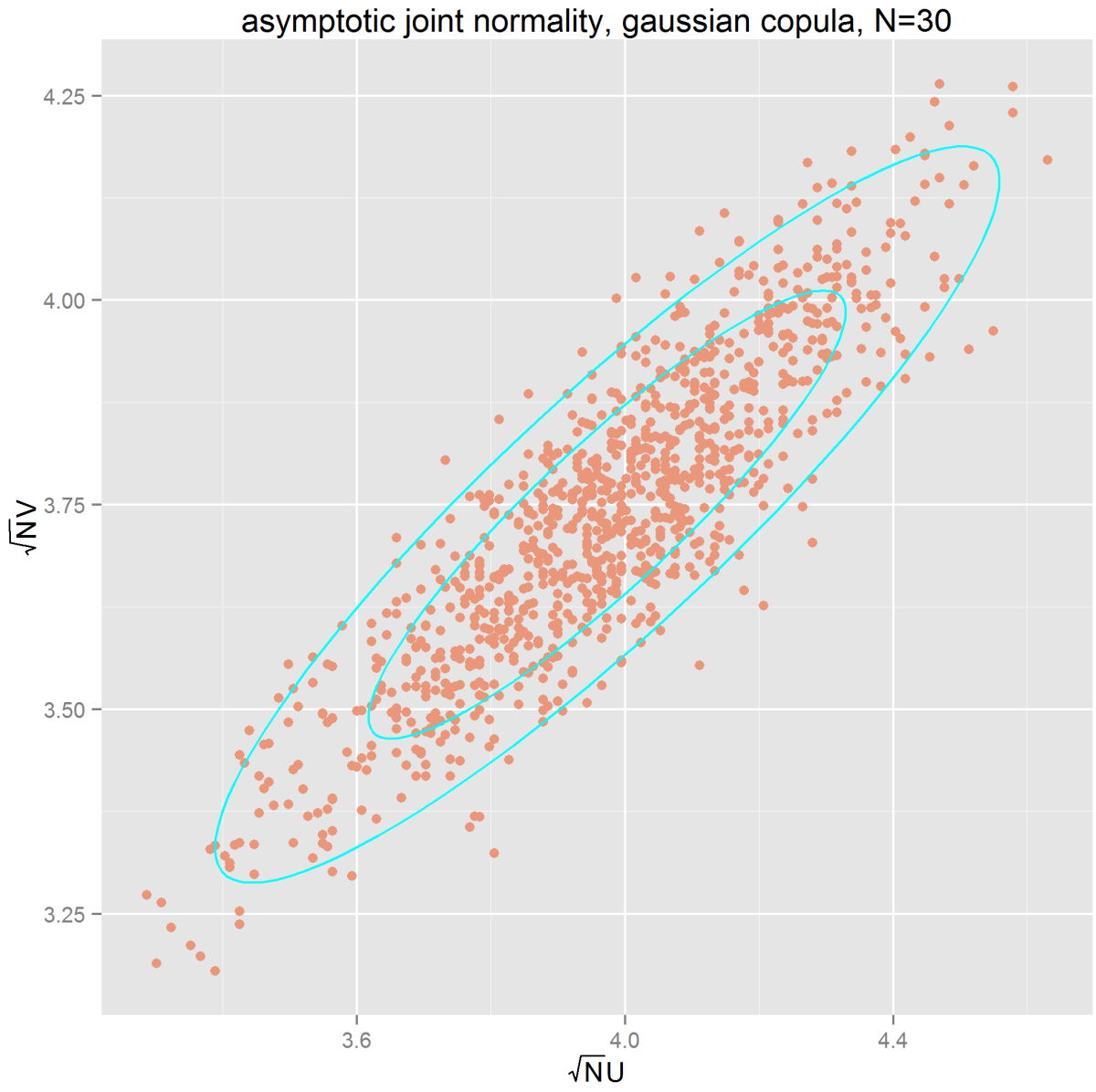

Figure 3: Scatterplot of 1e3 realizations of (U_N, V_N), along with 67% and 95% level curves of the $N((\hat{\theta_{11}}, \hat{\theta_{12}}), \hat{\Sigma})$ density. As before, the data in each replicate consists of $N = 30$ pairs (X_i, Y_i) of correlated vectors of length $n = m = 20$ with uniform marginals, obtained by applying the normal CDF to correlated normal vectors. The hats are meant to indicate that the quantities are empirical plug-in estimates of the asymptotic mean, variance, and covariance formulas given above.