Supporting Information for “Finite-Sample Two-Group Composite Hypothesis Testing via Machine Learning”

Tianyu Zhan and Jian Kang

1 Web Appendix A: sensitivity analyses for the scale-uniform distribution under different scenarios

In the main article Section 4.1 of the scale-uniform distribution, we consider $n = 20$, $k = 0.2, 0.8$, and $\theta_1 = 1, 5$. In this section, we provide additional simulation results under $n = 30, 40$, $k = 0.3, 0.5, 0.7$ and $\theta_1 = 2, 7$. Our DNN-based method has superior power as compared with alternatives under those scenarios (Web Table 1).
Web Table 1: Sensitivity analyses in the scale-uniform distribution under varying scenarios.

| $n$ | $k$ | $\theta_1$ | $\theta_2$ | Type I error rate (italicized) / power |
|-----|-----|-------------|-------------|--------------------------------------|
|     |     |             |             | DNN        | LRT        | Student’s $t$ | Wilcoxon |
| 30  | 0.5 | 2           | 2           | 4.9%       | 5.1%       | 5.0%          | 5.0%     |
|     |     |             |             | 2.181      | 92.3%      | 90.6%         | 30.5%    |
|     |     |             |             | 2.226      | 97.6%      | 96.4%         | 40.5%    |
|     |     |             |             | 2.249      | 98.6%      | 97.7%         | 45.8%    |
| 40  | 0.5 | 2           | 2           | 5.1%       | 5.0%       | 5.0%          | 5.0%     |
|     |     |             |             | 2.157      | 96.2%      | 94.9%         | 31.0%    |
|     |     |             |             | 2.196      | 99.0%      | 98.3%         | 41.2%    |
|     |     |             |             | 2.216      | 99.5%      | 99.0%         | 46.5%    |
| 20  | 0.7 | 2           | 2           | 5.0%       | 5.1%       | 5.0%          | 4.8%     |
|     |     |             |             | 2.311      | 86.5%      | 85.8%         | 28.7%    |
|     |     |             |             | 2.388      | 94.3%      | 93.8%         | 37.7%    |
|     |     |             |             | 2.427      | 96.4%      | 95.8%         | 42.4%    |
| 20  | 0.3 | 7           | 7           | 5.0%       | 5.0%       | 5.0%          | 4.8%     |
|     |     |             |             | 7.466      | 82.1%      | 72.8%         | 30.4%    |
|     |     |             |             | 7.582      | 92.7%      | 86.5%         | 40.6%    |
|     |     |             |             | 7.641      | 95.2%      | 90.5%         | 46.1%    |

2 Web Appendix B: sensitivity analyses for TS-DNN with different input data

Our method in the main article Section 4.1 utilizes sufficient statistics for $t^{(s)}$ when training TS-DNN. We perform additional analysis with key summary statistics (i.e., mean, median, standard deviation, sample quantiles) or order statistics as input data with results shown
in Web Table 2. All methods have relatively accurate type I error rate control. On power performance, TS-DNN with order statistics has smaller power than other two methods when $\theta = 1$, while has similar power when $\theta = 5$. Key summary statistics generally have similar performance with sufficient statistics, but is less powerful when $\theta = 1$ and $k = 0.2$. TS-DNN with sufficient statistics as implemented in the main text has the best performance in this problem.

3 Web Appendix C: sensitivity analyses for DNN performance under varying hyperparameters

In this section, we evaluate the performance of DNN with varying hyperparameters. Web Table 3 evaluates 6 different combinations of DNN structures for the first TS-DNN and the second CV-DNN. This first row corresponds to shallow neural networks with only 1 hidden layer, while the 6th row is for relatively deeper DNNs with 4 hidden layers. The 2nd to 5th rows contain DNN structures with moderate numbers of hidden layers at 2 and 3. Those structures are also evaluated as candidates in the cross-validation to select DNN structures in the main article. We observe that the performance of DNN is generally consistent, but deeper DNNs with larger numbers of layers tend to have higher power, for example the 5th row of TS-DNN with 2 layers and CV-DNN with 3 layers. The shallow neural networks with only 1 hidden layer in the 1st row has relatively lower power than deep neural networks. The reason is that DNN with more than 1 hidden layer usually has
Web Table 2: Sensitivity analyses in the scale-uniform distribution with different input data $t^{(s)}$ for TS-DNN.

| $k$ | $\theta_1$ | $\theta_2$ | Type I error rate (italicized) / power |
|-----|-------------|-------------|--------------------------------------|
|     |             |             | Sufficient statistics | Key summary statistics | Order statistics |
| 0.2 | 1           | 1           | 5.0%                    | 4.7%                   | 5.9%             |
|     | 1.044       | 76.2%       | 55.9%                   | 51.1%                  |
|     | 1.055       | 89.9%       | 72.5%                   | 67.0%                  |
|     | 1.061       | 93.7%       | 79.2%                   | 74.1%                  |
| 0.2 | 5           | 5           | 4.8%                    | 4.9%                   | 5.3%             |
|     | 5.222       | 80.2%       | 79.3%                   | 78.6%                  |
|     | 5.277       | 91.3%       | 91.7%                   | 91.3%                  |
|     | 5.305       | 94.5%       | 94.8%                   | 94.7%                  |
| 0.8 | 1           | 1           | 4.7%                    | 4.6%                   | 5.3%             |
|     | 1.178       | 87.5%       | 83.3%                   | 57.1%                  |
|     | 1.222       | 94.7%       | 93.0%                   | 72.7%                  |
|     | 1.244       | 96.6%       | 95.4%                   | 79.0%                  |
| 0.8 | 5           | 5           | 4.9%                    | 5.0%                   | 5.1%             |
|     | 5.888       | 88.0%       | 88.4%                   | 88.0%                  |
|     | 6.109       | 95.0%       | 95.2%                   | 95.0%                  |
|     | 6.220       | 96.8%       | 96.9%                   | 96.7%                  |

better generalization than shallow neural networks (Goodfellow et al., 2016). We suggest selecting a proper DNN structure by cross-validation as implemented in the main article.

Next in Web Table 4, we study the impact of batch size by fixing the structure for TS-DNN at 2 layers with 100 nodes per layer and CV-DNN at 2 layers with 50 nodes per layer with other hyperparameters same as the main article. All six combinations of batch sizes have similar power performance, while the first one with 10,000 for the TS-DNN and
10 for the CV-DNN has some numerical power gain. At Web Table 5, we also consider different dropout rates at 0.05, 0.2 and 0.4 in addition to 0.1 used in main article. The power of DNN method is consistent with different values of dropout rate in the training process, but is generally better with moderately small dropout rate of 0.1 and 0.05.

Next we implement our method with different ranges of $\theta$ ($\theta > 0$) as our parameter of interest and $k$ ($0 < k < 1$) as the nuisance design parameter when generating training data for TS-DNN and CV-DNN. Similar to Web Table 4, we validate our method when $k = 0.5$ and $\theta_1 = 2$. In Web Table 6, we consider the following three sets of $\Theta$ as the range for $\theta$, and $K$ as the range for $k$:

1. $\Theta = (0.5, 10)$, $K = (0, 1)$;
2. $\Theta = (0.1, 15)$, $K = (0, 1)$;
3. $\Theta = (4, 7)$, $K = (0, 0.3),$

where Range 1 is utilized in the main article, Range 2 is wider, while Range 3 is narrower. Based on Web Table 6, we observe that DNN has consistent performance when training data are sampled from Range 1 and 2. However, the type I error rate is not controlled at $\alpha = 5\%$ under Range 3 because the validation values of $\theta$ and $k$ are out of range. Therefore, it is important to make sure that the ranges of training data are wide enough to cover the validation values when implementing our method. Moreover, the training data size $A \times (B_0 + B_1)$ for TS-DNN and $A$ for CV-DNN can be set sufficiently large to accommodate wider ranges and higher dimensions of simulated underlying parameters. To further demonstrate the robustness of our method, we replicate the results of DNN in main
Finally, we consider the *softmax* function as the last-layer activation function in TS-DNN training, as an alternative to the *sigmoid* function used in the main article. Based on Web Table 7, both activation functions have similar power performance with controlled type I error rate. The *sigmoid* activation function has slightly higher power when $k = 0.2$ and $\theta = 1$, while the *softmax* is more powerful in other scenarios.

The overall conclusion is that the performance our method in terms of controlling type I error rate and gaining power is consistent with different properly chosen hyperparameters in DNN training. One can also perform grid-search optimization based on cross-valuation to find the best set of parameters for a particular problem.

### 4 Web Appendix D: sensitivity analyses with other machine learning methods

We further evaluate the performance of three alternatives discussed in main text Section 3.3 by substituting DNN with other approaches in the two-stage method of Figure 1: GLM/LM, SVM and RF with the same computational cost. GLM/LM utilizes generalized linear model (GLM) in the first stage to approximate test statistics and linear model (LM) in the second stage to estimate critical values. We evaluate three SVMs with different kernel types: SVM(p) with polynomial and degree of 3, SVM(s) with sigmoid and SVM(r) with radial. Other hyperparameters are set at their default values in the R package e1071.
(Meyer et al., 2019). We evaluate two RF methods: RF(500) with 500 trees (default setting) and RF(1,000) with 1,000 trees. Other parameters are kept at their default values in the R package randomForest (Liaw and Wiener, 2002). We use the training data setting of DNN in the main text Section 4.1 as a benchmark for the computational cost. The first stage training data size is $10^7$ ($B_0 = B_1 = 10^4$ samples per feature and $A = 500$ parameter features), and the second stage training data size is $A = 500$ with $B' = 10^6$ simulated null data per feature. With the same computation time, SVM needs to reduce its training data size to $B_0 = B_1 = 150$ and $B' = 30,000$. RF(500) reduces to $B_0 = B_1 = 150$, and RF(1,000) reduces to $B_0 = B_1 = 100$, but $B'$ stays the same with DNN at $10^6$. Additionally, we consider a DNN(r) method with reduced training features at $B_0 = B_1 = 100$ and $B' = 30,000$ to match the worst setting in both SVM and RF. The number of training epochs for the TS-DNN is increased from 10 to 100 to accommodate the smaller data size.

As shown in Web Table 8, our proposed DNN method with the benchmark setting has accurate error rates at $\alpha = 5\%$ across all scenarios. Under the same computational cost, GLM/LM, SVM(p), SVM(s), SVM(r), RF(500) and RF(1,000) have inflated or deflated type I error rates in some scenarios. Even with reduced training features, DNN(r) still has a satisfactory type I error rate control. This evidence of accommodating a larger training data size also supports DNN’s scalability on large datasets.
5 Web Appendix E: saving of computational time

We mention in main text Section 3.3 that another advantage of our proposed method is the saving on computational time. As shown in Web Table 9, it takes 0.5 hour to obtain the test statistics from the TS-DNN in Section 3.2. The proposed two-stage method then spends approximately 3 hours to generate \( A = 500 \) training data and then fit the CV-DNN in Section 3.3. Therefore, its computational time of 1.02 hour in the validation stage is much shorter than the over 90,000 hours from the parametric bootstrap method, which has to simulate \( B' = 10^6 \) null data in each of the \( 10^6 \) validation iterations and 16 scenarios.

6 Web Appendix F: sensitivity analyses for the ACTT

In the main text Section 5.1 on ACTT, we consider an adaptive design with \( n_{min}^{(2)} = 30, n_{max}^{(2)} = 400 \), and \( \theta_{min} = 0.1 \). We conduct additional simulations in Web Table 10 with varying magnitudes of parameters. The first and the second blocks evaluate different magnitudes of \( \theta_1 \), while the third and the forth blocks consider varying \( n_{min}^{(2)} \) and \( n_{max}^{(2)} \). The DNN approach has superior performance as compared with two alternatives, INCTA and ET, across all scenarios.

References

Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio (2016). *Deep learning*, Volume 1. MIT press Cambridge.
Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News* 2(3), 18–22.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. *R package version 1.7-3*. https://CRAN.R-project.org/package=e1071.
Web Table 3: Sensitivity analyses in the scale-uniform distribution under varying DNN structures.

| Structure parameters | n  | k  | $\theta_1$ | $\theta_2$ | Type I error rate (italicized) / power |
|----------------------|----|----|------------|------------|---------------------------------------|
| TS-DNN$^1$ CV-DNN$^1$|    |    |            |            |                                       |
| 1 / 100              | 1  | 50 | 20         | 0.5        | 2                                    | 5.0%   |
|                      |    |    |            |            |                                       | 2.222  |
|                      |    |    |            |            |                                       | 2.277  |
|                      |    |    |            |            |                                       | 2.305  |
| 2 / 50               | 2  | 100| 20         | 0.5        | 2                                    | 4.9%   |
|                      |    |    |            |            |                                       | 2.222  |
|                      |    |    |            |            |                                       | 2.277  |
|                      |    |    |            |            |                                       | 2.305  |
| 3 / 150              | 2  | 100| 20         | 0.5        | 2                                    | 5.1%   |
|                      |    |    |            |            |                                       | 2.222  |
|                      |    |    |            |            |                                       | 2.277  |
|                      |    |    |            |            |                                       | 2.305  |
| 2 / 100              | 2  | 50 | 20         | 0.5        | 2                                    | 5.0%   |
|                      |    |    |            |            |                                       | 2.222  |
|                      |    |    |            |            |                                       | 2.277  |
|                      |    |    |            |            |                                       | 2.305  |
| 2 / 100              | 3  | 150| 20         | 0.5        | 2                                    | 5.1%   |
|                      |    |    |            |            |                                       | 2.222  |
|                      |    |    |            |            |                                       | 2.277  |
|                      |    |    |            |            |                                       | 2.305  |
| 4 / 100              | 4  | 50 | 20         | 0.5        | 2                                    | 5.0%   |
|                      |    |    |            |            |                                       | 2.222  |
|                      |    |    |            |            |                                       | 2.277  |
|                      |    |    |            |            |                                       | 2.305  |

$^1$ The number of hidden layers / the number of nodes per layer.
Web Table 4: Sensitivity analyses in the scale-uniform distribution under varying DNN batch sizes.

| Batch size | n     | k   | $\theta_1$ | $\theta_2$ | Type I error rate (italicized) / power |
|------------|-------|-----|-------------|-------------|----------------------------------------|
|            |       |     |             |             | TS-DNN | CV-DNN | DNN        |                     |
| 10,000     | 10    | 20  | 0.5         | 2           | 2     | 4.9%    | 2.222      | 84.4%                |
|            |       |     |             |             |       |         | 2.277      | 93.4%                |
|            |       |     |             |             |       |         | 2.305      | 95.7%                |
| 1,000      | 10    | 20  | 0.5         | 2           | 2     | 4.9%    | 2.222      | 83.8%                |
|            |       |     |             |             |       |         | 2.277      | 93.6%                |
|            |       |     |             |             |       |         | 2.305      | 96.0%                |
| 100        | 10    | 20  | 0.5         | 2           | 2.000 | 5.0%    | 2.222      | 84.0%                |
|            |       |     |             |             |       |         | 2.277      | 93.5%                |
|            |       |     |             |             |       |         | 2.305      | 95.8%                |
| 10,000     | 100   | 20  | 0.5         | 2           | 2     | 4.9%    | 2.222      | 82.6%                |
|            |       |     |             |             |       |         | 2.277      | 92.4%                |
|            |       |     |             |             |       |         | 2.305      | 94.9%                |
| 1,000      | 100   | 20  | 0.5         | 2           | 2     | 5.0%    | 2.222      | 84.6%                |
|            |       |     |             |             |       |         | 2.277      | 93.7%                |
|            |       |     |             |             |       |         | 2.305      | 95.9%                |
| 100        | 100   | 20  | 0.5         | 2           | 2.000 | 5.1%    | 2.222      | 84.0%                |
|            |       |     |             |             |       |         | 2.277      | 93.3%                |
|            |       |     |             |             |       |         | 2.305      | 95.7%                |
Web Table 5: Sensitivity analyses in the scale-uniform distribution under varying DNN dropout rates.

| Dropout rate | n  | k  | $\theta_1$ | $\theta_2$ | Type I error rate (*italicized*) / power of DNN |
|--------------|----|----|------------|------------|-----------------------------------------------|
| 0.1          | 20 | 0.5| 2          | 2          | 4.9%                                           |
|              |    |    | 2.222      | 84.4%      |                                               |
|              |    |    | 2.277      | 93.4%      |                                               |
|              |    |    | 2.305      | 95.7%      |                                               |
| 0.05         | 20 | 0.5| 2          | 2          | 4.9%                                           |
|              |    |    | 2.222      | 84.6%      |                                               |
|              |    |    | 2.277      | 93.9%      |                                               |
|              |    |    | 2.305      | 96.1%      |                                               |
| 0.2          | 20 | 0.5| 2          | 2          | 5.1%                                           |
|              |    |    | 2.222      | 84.5%      |                                               |
|              |    |    | 2.277      | 93.7%      |                                               |
|              |    |    | 2.305      | 95.9%      |                                               |
| 0.4          | 20 | 0.5| 2          | 2          | 5.1%                                           |
|              |    |    | 2.222      | 83.6%      |                                               |
|              |    |    | 2.277      | 93.3%      |                                               |
|              |    |    | 2.305      | 95.7%      |                                               |
Web Table 6: Sensitivity analyses in the scale-uniform distribution under varying ranges when simulating training data for DNN.

| Range       | n  | k  | θ₁ | θ₂ | Type I error rate (italicized) / power |
|-------------|----|----|----|----|---------------------------------------|
| (0, 1)      | 20 | 0.5| 2  | 2  | 4.9%                                  |
|             |    |    |    |    | 2.222                                 |
|             |    |    |    |    | 2.277                                 |
|             |    |    |    |    | 2.305                                 |
| (0, 1)      | 20 | 0.5| 2  | 2  | 4.9%                                  |
|             |    |    |    |    | 2.222                                 |
|             |    |    |    |    | 2.277                                 |
|             |    |    |    |    | 2.305                                 |
| (0, 0.3)    | 20 | 0.5| 2  | 2  | 100.0%                                |
|             |    |    |    |    | 2.222                                 |
|             |    |    |    |    | 2.277                                 |
|             |    |    |    |    | 2.305                                 |
Web Table 7: Sensitivity analyses in the scale-uniform distribution with different last-layer activation function for TS-DNN

|   |   |   | Type I error rate (italicized) / power |
|---|---|---|---------------------------------------|
|   |   |   | **sigmoid** | **softmax** |
| 0.2 | 1 | 1 | 5.0% | 5.2% |
|     | 1.044 | | 76.2% | 73.5% |
|     | 1.055 | | 89.9% | 87.4% |
|     | 1.061 | | 93.7% | 91.5% |
| 0.2 | 5 | 5 | 4.8% | 5.1% |
|     | 5.222 | | 80.2% | 81.4% |
|     | 5.277 | | 91.3% | 92.3% |
|     | 5.305 | | 94.5% | 95.2% |
| 0.8 | 1 | 1 | 4.7% | 4.9% |
|     | 1.178 | | 87.5% | 88.0% |
|     | 1.222 | | 94.7% | 94.9% |
|     | 1.244 | | 96.6% | 96.6% |
| 0.8 | 5 | 5 | 4.9% | 4.9% |
|     | 5.888 | | 88.0% | 88.4% |
|     | 6.109 | | 95.0% | 95.3% |
|     | 6.220 | | 96.8% | 96.9% |
Web Table 8: Type I error rates of four methods in the proposed two-stage method under a nominal level of 5% with the same computational cost.

| $k$ | $\theta_1 = \theta_2$ | With the same computational cost | | | | | |
|-----|------------------------|---------------------------------|-----|-----|-----|-----|-----|
|     |                        | DNN$^1$ | GLM/LM$^2$ | SVM(p)$^3$ | SVM(s)$^4$ | SVM(r)$^5$ | RF(500)$^6$ | RF(1,000)$^7$ | DNN(r)$^8$ |
| 0.2 | 1                      | 5.0%   | 20.5%    | 0.0%      | 0.0%      | 1.3%      | 38.8%      | 34.4%      | 4.9%      |
|     | 5                      | 4.8%   | 5.0%     | 6.7%      | 100.0%    | 5.3%      | 3.2%       | 2.9%       | 4.7%      |
| 0.8 | 1                      | 4.7%   | 3.4%     | 10.9%     | 100.0%    | 3.7%      | 30.6%      | 30.8%      | 4.6%      |
|     | 5                      | 4.9%   | 5.0%     | 5.1%      | 0.0%      | 5.0%      | 11.0%      | 16.2%      | 4.7%      |

1. DNN in the two stages with $B_0 = B_1 = 10,000$, $B' = 1,000,000$, and the training epoch of TS-DNN at 10.
2. GLM in the first stage and LM in the second stage with $B_0 = B_1 = 10,000$ and $B' = 1,000,000$.
3. SVM with polynomial kernel of degree 3 in the two stages with $B_0 = B_1 = 150$ and $B' = 30,000$.
4. SVM with sigmoid kernel in the two stages with $B_0 = B_1 = 150$ and $B' = 30,000$.
5. SVM with radial kernel in the two stages with $B_0 = B_1 = 150$ and $B' = 30,000$.
6. RF with 500 trees in the two stages with $B_0 = B_1 = 150$ and $B' = 1,000,000$.
7. RF with 1,000 trees in the two stages with $B_0 = B_1 = 100$ and $B' = 1,000,000$.
8. DNN in the two stages with $B_0 = B_1 = 100$, $B' = 30,000$, and the training epoch of TS-DNN at 100.

Web Table 9: Two-stage DNN-based method saves considerable computational time (in hours) as compared with parametric bootstrap method.

| Method                        | TS-DNN | CV-DNN | Validation | Total time |
|-------------------------------|--------|--------|------------|------------|
| Two-stage DNN-based method    | 0.50   | 2.95   | 1.02       | 4.47       |
| TS-DNN + parametric bootstrap | 0.50   | -      | >90,000    | >90,000    |
Web Table 10: Sensitivity analyses in the ACTT on COVID-19 under varying scenarios.

| $n_{\text{min}}$ | $n_{\text{max}}$ | $\theta_{\text{min}}$ | $\theta_1$ | $\theta_2$ | Type I error rate (*italicized*) / power |
|------------------|-----------------|-----------------|----------|----------|--------------------------------------|
|                  |                 |                 |          |          | DNN   | INCTA\(^1\) | ET   |
| 30               | 400             | 0.1             | 0.42     | 0.42     | 0.53  | 89.2%    | 83.8% | 87.6%\(^2\) |
|                  |                 |                 |          |          | 0.54  | 93.9%    | 86.9% | 90.8%\(^2\) |
|                  |                 |                 |          |          | 0.55  | 96.5%    | 89.1% | 92.9%\(^2\) |
| 30               | 400             | 0.1             | 0.52     | 0.52     | 0.63  | 90.6%    | 85.3% | 88.6%\(^3\) |
|                  |                 |                 |          |          | 0.64  | 94.9%    | 88.4% | 91.7%\(^3\) |
|                  |                 |                 |          |          | 0.65  | 97.2%    | 90.4% | 93.6%\(^3\) |
| 50               | 300             | 0.12            | 0.47     | 0.47     | 0.58  | 90.9%    | 87.8% | 89.6%\(^4\) |
|                  |                 |                 |          |          | 0.59  | 95.2%    | 92.0% | 93.9%\(^4\) |
|                  |                 |                 |          |          | 0.60  | 97.6%    | 94.7% | 96.5%\(^4\) |
| 20               | 450             | 0.12            | 0.47     | 0.47     | 0.58  | 96.1%    | 88.3% | 94.0%\(^5\) |
|                  |                 |                 |          |          | 0.59  | 98.5%    | 90.3% | 96.1%\(^5\) |
|                  |                 |                 |          |          | 0.60  | 99.5%    | 91.5% | 97.2%\(^5\) |

\(^1\) The inverse normal combination test approach (INCTA).
\(^2\) The empirical test (ET) based on a critical value of 0.033 in the $p$-value scale.
\(^3\) ET based on a critical value of 0.033 in the $p$-value scale.
\(^4\) ET based on a critical value of 0.041 in the $p$-value scale.
\(^5\) ET based on a critical value of 0.036 in the $p$-value scale.