Simulation with different distances

Jiaxin Hu

May 8, 2022

We run a small scale simulation to verify the accuracy of Gaussian matrix matching algorithms with different distance statistics.

1 Introduction

The observations $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^n$ comes from the Gaussian Winger model; i.e.,

$$(\boldsymbol{A}_{ij}, \boldsymbol{B}'_{ij}) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad (\boldsymbol{A}_{ij}, \boldsymbol{B}'_{ij}) \perp (\boldsymbol{A}_{i'j'}, \boldsymbol{B}'_{i'j'}) \quad \text{for all } (i,j) \neq (i',j'),$$

and

$$\boldsymbol{B} = \Pi^* \boldsymbol{B}' (\Pi^*)^T,$$

where $\Pi^* \in \{0,1\}^{n \times n}$ is the permutation matrix with $\Pi^*_{ij} = \mathbb{1}\{j = \pi^*(i)\}$ and $\pi^* : [n] \mapsto [n]$ is the true permutation on $[n]$. Let $\sigma = \sqrt{1-\rho^2}$.

To evaluate the matching accuracy, we consider the ratio of node pair agreement between the estimated permutation $\hat{\pi}$ and true permutation

$$OV_n(\hat{\pi}, \pi^*) = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{\hat{\pi}(i) \neq \pi^*(i)\}.$$

Consider the observations $\boldsymbol{A}, \boldsymbol{B}$ comes from the Gaussian Winger model and an arbitrary pair $(a,b) \in [n]^2$. For simplicity, we rename the samples as $\{X_i\}_{i\in[n]} \coloneqq \{\boldsymbol{A}_{ai}\}_{i\in[n]}$ and $\{Y_i\}_{i\in[n]} \coloneqq \{\boldsymbol{B}_{bi}\}_{i\in[n]}$. Let $f_n = \frac{1}{n}\sum_{i\in[n]} \delta_{X_i}$ and $g_n = \frac{1}{n}\sum_{i\in[n]} \delta_{Y_i}$ denote the empirical PDFs, and let $\hat{F}_n(t) = \frac{1}{n}\sum_{i\in[n]} \mathbb{1}\{X_i \leq t\}$ $\hat{G}_n(t) = \frac{1}{n}\sum_{i\in[n]} \mathbb{1}\{Y_i \leq t\}$ denote the empirical CDFs. Let $B_1 = \min\{\min_{i\in[n]}\{X_i\}, \min_{i\in[n]}\{Y_i\}\}$ and $B_2 = \max\{\max_{i\in[n]}\{X_i\}, \max_{i\in[n]}\{Y_i\}\}$.

We now consider 8 distance statistics for $\{X_i\}, \{Y_i\}$.

1. **Smoothed TV norm $\hat{T}V(X,Y|L)$.** Analogy of the Z distance proposed in Ding et al. (2021).

   $$\hat{T}V(X,Y|L) = \sum_{l\in[L]} |f_n(I_l) - g_n(I_l)|\,,$$

   where $\{I_l\}_{l\in[L]}$ is the uniform partition over the interval $[-1/2, 1/2]$.

2. **Smoothed W_1 norm $\hat{W}_1(X,Y|L)$.** Consider the uniform partition $\{I_l\}_{l\in[L]}$ over the interval $[-1/2,1/2]$. We define

$$\hat{W}_1(X,Y|L) = \sum_{l\in[L]} |F_n(t_l) - G_n(t_l)|,$$

where t_l's are the right boundaries of I_l's.

3. **Smoothed W_1 norm $\hat{W}_1(X,Y|2n)$.** Defined as $\hat{W}_1(X,Y|L)$ with specified $L=2n$.

4. **Smoothed W_1 norm with different partition $\tilde{W}_1(X,Y|L)$.** Consider the uniform partition $\{I_l\}_{l\in[L]}$ over the interval $[-L/2,L/2]$. Then define

$$\tilde{W}_1(X,Y|L) = \sum_{l\in[L]} |F_n(t_l) - G_n(t_l)|,$$

where t_l's are the right boundaries of I_l's.

5. **Smoothed W_1 norm with different partition $\tilde{W}_1(X,Y|2n)$.** Defined as $\tilde{W}_1(X,Y|L)$ with specified $L=2n$.

6. **Smoothed W_1 norm with unfixed-boundary partition $\bar{W}_1(X,Y|L)$.** Consider the $2n$-uniform partition $\{I_l\}_{l\in[L]}$ over the interval $[B_1,B_2]$. Then define

$$\bar{W}_1(X,Y|L) = \sum_{l\in[L]} |F_n(t_l) - G_n(t_l)|,$$

where t_l's are the right boundaries of I_l's.

7. **Smoothed W_1 norm with unfixed-boundary partition $\bar{W}_1(X,Y|2n)$.** Defined as $\bar{W}_1(X,Y|L)$ with specified $L=2n$.

8. **Empirical W_1 norm $\hat{W}_1(X,Y)$.** Let $X_{(i)}$ and $Y_{(i)}$ denote the order statistics of $\{X_i\}$ and $\{Y_i\}$. We define

$$\hat{W}_1(X,Y) = \int_{\mathbb{R}} |F_n(t) - G_n(t)|dt = \int_0^1 |F_n^{-1}(u) - G_n^{-1}(u)|du = \frac{1}{n}\sum_{i\in[n]} |X_{(i)} - Y_{(i)}|.$$

Rename the observations $X_1,\dots,X_n,Y_1,\dots,Y_n$ as $U_1,\dots,U_{2n}$ with order statistics $U_{(i)}$. The empirical W_1 is equal to

$$\hat{W}_1(X,Y) = \sum_{k=2}^{2n} |F_n(U_{(k)}) - G_n(U_{(k)})| \cdot |U_{(k)} - U_{(k-1)}|.$$

We summarize the differences between different distances:

1. $\hat{TV}(X,Y|L)$ vs $\hat{W}_1(X,Y|L)$. Both distances consider L-uniform partition over $[-1/2,1/2]$ with window size $1/L$. But $\hat{TV}$ uses smoothed PDF while $\hat{W}_1$ uses smoothed CDF.

2. $\hat{W}_1(X,Y|L)$ vs $\hat{W}_1(X,Y|2n)$. The latter one has the same definition as former one with specified $L=2n$.

3. $\hat{W}_1(X,Y|L)$ vs $\tilde{W}_1(X,Y|L)$. Both consider the L-uniform partition and smoothed CDF. But the former one uses the partition over $[-1/2, 1/2]$ with window size $1/L$ while the latter one uses partition over $[-L/2, L/2]$ with window size 1.

4. $\tilde{W}_1(X,Y|L)$ vs $\tilde{W}_1(X,Y|2n)$. The latter one has the same definition as former one with specified $L = 2n$.

5. $\tilde{W}_1(X,Y|L)$ vs $\bar{W}_1(X,Y|L)$. Both consider L-uniform partition. But former one considers partition over $[-L/2, L/2]$ and the latter one uses partition only cover the minima and maxima of all observations, $[B_1, B_2]$.

6. $\bar{W}_1(X,Y|L)$ vs $\bar{W}_1(X,Y|2n)$. The latter one has the same definition as former one with specified $L = 2n$.

7. $\bar{W}_1(X,Y|2n)$ vs $\hat{W}_1(X,Y)$. Both consider the $2n$ partition over $[B_1, B_2]$. But the former one considers the uniform partition with fixed window size $(B_2 - B_1)/2n$ while the latter one considers the unfixed window size $|U_{(i+1)} - U_{(i)}|$, where $U_{(i)}$'s for $i \in [2n]$ are the order statistics for the mixed observations $X_1, \ldots, X_n, Y_1, \ldots, Y_n$.

2 Simulation result

We consider the set up $n = 50$ with varying $\sigma \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. A smaller σ indicates a larger correlation between paired edges. For distances relied on L, $\hat{TV}(X,Y|L)$, $\hat{W}_1(X,Y|L)$, $\tilde{W}_1(X,Y|L)$, and $\bar{W}_1(X,Y|L)$, we try L from 5 to $3n$ with gap 5 (i.e., $L \in \{5, 10, \ldots, 3n\}$) and pick the best result. Figure 1 shows the simulation results with 7 different methods.

Here are a few observations from Figure 1.

1. Empirical PDF may not work as good as empirical CDF in practice. Notice that smoothed TV norm (black solid line) is worse than all other CDF-based methods. Particularly, the smoothed W_1 (blue solid line) uses the same partition as smoothed TV and outperforms the TV norm.

2. The choice of L may not lead to dramatic accuracy changes. But $L = 2n$ is not always the optimal choice. Note that the smoothed W_1 with $L = 2n$ (blue, green, red dashed lines and pink solid line) are lower than that with optimal L via exhausting search (blue, green, and red solid lines).

3. The range of partition may be the critical factor for the accuracy. The $\hat{W}_1$'s with L (blue lines) using partition covers $[-1/2, 1/2]$, the $\tilde{W}_1$'s with L (green lines) using partition covers $[-L/2, L/2]$, and $\bar{W}_1, \hat{W}_1$ (red lines) using partition covers $[B_1, B_2]$ which is exactly the range of the observations X_i, Y_i.

4. Window size of the partition may not be the critical factor. Note that the only difference between $\bar{W}_1$ (red lines) and $\hat{W}_1$ (pink solid line) is the window size. But their performances are very similar.

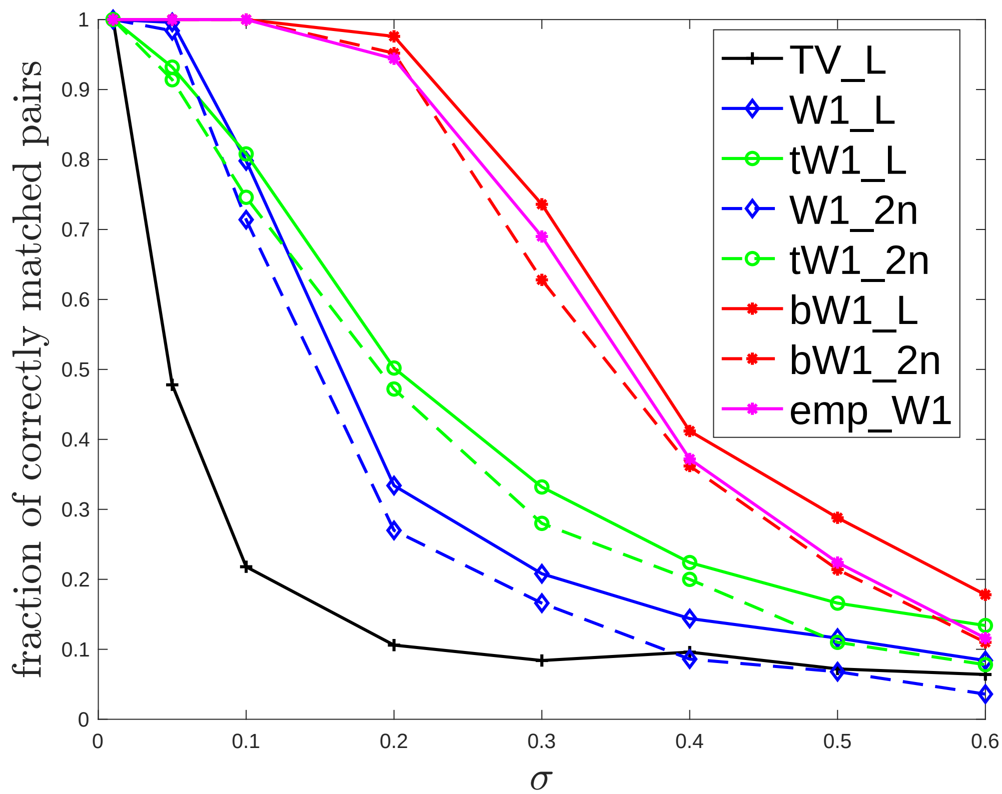

Figure 1: The ratio of correct agreement $OV_n(\hat{\pi}, \pi^*)$ versus the noise level σ with 7 different methods: $\hat{TV}(X, Y|L)$ **(TV_L)**, $\hat{W}_1(X, Y|L)$ **(W1_L)**, $\hat{W}_1(X, Y|2n)$ **(W1_2n)**, $\tilde{W}_1(X, Y|L)$ **(tW1_L)**, $\tilde{W}_1(X, Y|2n)$ **(tW1_2n)**, $\bar{W}_1(X, Y|L)$ **(bW1_L)**, $\bar{W}_1(X, Y|2n)$ **(bW1_2n)**, and $\hat{W}_1(X, Y)$ **(emp_W1)**.

References

Ding, J., Ma, Z., Wu, Y., and Xu, J. (2021). Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1):29–115.