Inference for Projection-Based Wasserstein Distances on Finite spaces

Ryo Okano[1] and Masaaki Imaizumi[1,2]

[1] *The University of Tokyo,* [2] *RIKEN Center for AIP*

Supplementary Material

This supplement material contains a part of the proofs of the theorems, propositions and lemmas given in the main text, as well as additional simulation results.

S1 Proofs

S1.1 Proof of proposition 1

Let $\{h_{1\ell}\}, \{h_{2\ell}\} \subset \Omega_N$ be sequences satisfying $h_{1\ell} \to h_1, h_{2\ell} \to h_2$, and let $t_\ell \searrow 0$ as $\ell \to \infty$. Following the definition of directional Hadamard derivative, we consider the following difference:

$$\begin{aligned}&\frac{\mathrm{IW}_p^p(r+t_\ell h_{1\ell}, s+t_\ell h_{2\ell}) - \mathrm{IW}_p^p(r,s)}{t_\ell}.\\ &= \int_{S_{d,k}} \frac{W_p^p(r+t_\ell h_{1\ell}, s+t_\ell h_{2\ell}; \mathcal{X}_E) - W_p^p(r,s;\mathcal{X}_E)}{t_\ell} d\mu(E),\end{aligned} \tag{S1.1}$$

and consider its limit. For each $E \in S_{d,k}$, Theorem 4 in Sommerfeld and Munk (2018) implies

$$\frac{W_p^p(r+t_\ell h_{1\ell}, s+t_\ell h_{2\ell}; \mathcal{X}_E) - W_p^p(r,s;\mathcal{X}_E)}{t_\ell} \to \max_{(u,v)\in\Phi_p^*(r,s;\mathcal{X}_E)} -(\langle u, h_1\rangle + \langle v, h_2\rangle),$$

as $\ell \to \infty$. Furthermore, the Lipschitz continuity of the Wasserstein distance (Theorem 4 of Sommerfeld and Munk (2018)) implies

$$\begin{aligned}\left|\frac{W_p^p(r+t_\ell h_{1\ell}, s+t_\ell h_{2\ell}; \mathcal{X}_E) - W_p^p(r,s;\mathcal{X}_E)}{t_\ell}\right| &\leq \frac{p\mathrm{diam}(\mathcal{X}_E)^p \|t_\ell(h_{1\ell}, h_{2\ell})\|}{t_\ell} \\ &\leq pk^p \mathrm{diam}(\mathcal{X})^p \|(h_{1\ell}, h_{2\ell})\|.\end{aligned}$$

The last inequality results from $\mathrm{diam}(\mathcal{X}_E) \leq k\mathrm{diam}(\mathcal{X})$, which follows from

$$\|E^\top x\| \leq |E_1^\top x| + \cdots + |E_k^\top x| \leq \|E_1\|\|x\| + \cdots + \|E_k\|\|x\| = k\|x\|$$

for $E = (E_1, ..., E_k) \in S_{d,k}$ and $x \in \mathbb{R}^d$. Because $\mathcal{X}$ is finite and $\{h_{1\ell}\}$ and $\{h_{2\ell}\}$ are convergent sequences, $pk^p\mathrm{diam}(\mathcal{X})^p\|(h_{1\ell}, h_{2\ell})\|$ is bounded by a constant not depending on E and ℓ. Therefore, by taking $\ell \to \infty$ in (S1.1), we can apply the dominated convergence theorem, and the claim then holds.

S1.2 Proof of Theorem 2

Our proof follows the same line as the proof of Theorem 1 of Sommerfeld and Munk (2018). Under the assumption of the theorem, the central limit theorem implies

$$\sqrt{\frac{nm}{n+m}}\{(\hat{r}_n, \hat{s}_m) - (r,s)\} \xrightarrow{d} (\sqrt{\delta}G, \sqrt{1-\delta}H),$$

as $n \wedge m \to \infty$.

About (i): An application of the delta method in Theorem 1 with the directional Hadamard derivative of the map $(r,s) \mapsto \mathrm{IW}_p^p(r,s)$, which is given in Proposition 1, yields

$$\sqrt{\frac{nm}{n+m}}\mathrm{IW}_p^p(\hat{r}_n,\hat{s}_m) \xrightarrow{d} \int_{S_{d,k}} \max_{(u,v)\in\Phi_p^*(r,s;\mathcal{X}_E)} -(\langle u,\sqrt{\delta}G\rangle+\langle v,\sqrt{1-\delta}H\rangle)d\mu(E). \tag{S1.2}$$

Note that, under $r = s$, we have $(u,v) \in \Phi^*(r,s;\mathcal{X}_E)$ if and only if $u \in \Phi_p^*(\mathcal{X}_E)$ and $v = -u$. Therefore, with $G \overset{d}{=} H, -G \overset{d}{=} G$ and $-H \overset{d}{=} H$, we have

$$\begin{aligned}
\max_{(u,v)\in\Phi_p^*(r,s;\mathcal{X}_E)} -(\langle u,\sqrt{\delta}G\rangle + \langle v,\sqrt{1-\delta}H\rangle) &\overset{d}{=} \max_{(u,v)\in\Phi_p^*(\mathcal{X}_E)} \sqrt{\delta}\langle G,u\rangle - \sqrt{1-\delta}\langle H,u\rangle \\
&\overset{d}{=} \max_{(u,v)\in\Phi_p^*(\mathcal{X}_E)} \sqrt{\delta+(1-\delta)}\langle G,u\rangle \\
&= \max_{(u,v)\in\Phi_p^*(\mathcal{X}_E)} \langle G,u\rangle, \qquad \text{(S1.3)}
\end{aligned}$$

for each $E \in S_{d,k}$. (S1.2), (S1.3), and an application of the continuous mapping theorem with the map $t \mapsto t^{1/p}$ provides the conclusion.

About (ii): By Proposition 1 and the chain rule for directional Hadamard derivatives (Proposition 3.6 of Shapiro (1990)), the directional Hadamard derivative of the map $(r,s) \mapsto \mathrm{IW}_p(r,s) = (\mathrm{IW}_p^p(r,s))^{1/p}$ at (r,s) is given by

$$(h_1, h_2) \mapsto \frac{1}{p}\mathrm{IW}_p^{1-p}(r, s) \int_{S_{d,k}} \max_{(u,v)\in\Phi_p^*(r,s;\mathcal{X}_E)} -(\langle u, h_1\rangle + \langle v, h_2\rangle d\mu(E)).$$

An application of the delta method in Theorem 1 yields the conclusion.

Remark 1. In the proof of Theorem 2, we applied the delta method with IW_p^p when $r = s$, and IW_p when $r \neq s$, respectively. There are mainly two reasons for this usage. First, $\mathrm{IW}_p(r, s)$ is not directionally Hadamard differentiable when $r = s$. Second, applying the delta method to IW_p^p and using the map $t \mapsto t^{1/p}$ would not result in the correct scaling when $r \neq s$. To avoid these issues, we use the delta method differently for each of these situations.

S1.3 Proof of Lemma 1

Our proof is similar to the proof of Theorem 2.3 in Klatt *et al.* (2020), which shows continuous differentiability of the regularized optimal transport plan without projection. Note that the regularized optimal transport problem (2.4) with marginal r_0 and s_0 satisfies the Slater's constraint qualification (Proposition 26.18 in Bauschke *et al.* (2011)). Therefore, strong duality holds and the dual problem admits an optimal solution. In addition, we can characterize the regularized optimal transport plan $\pi_{p,\lambda}$ and its corresponding optimal dual solution $\mu_{p,\lambda} \in \mathbb{R}^{2N-1}$ by the necessary and sufficient

Karush-Kuhn-Tucker conditions:

$$c_p(\mathcal{X}_E) + \lambda \nabla \phi(\pi_{p,\lambda})^\top - A_\star^\top \mu_{p,\lambda} = 0, \quad A_\star \pi_{p,\lambda} - (r_0, s_{0\star})^\top = 0.$$

We now obtain the statement by applying the implicit function theorem to this system of equations. Let us define a function $F : \mathbb{R}^{N^2} \times \mathbb{R}^{2N-1} \times \mathbb{R}^{2N-1} \times \mathbb{R}^{dk} \to \mathbb{R}^{N^2+2N-1}$ by

$$F(\pi, \mu, (r, s_\star), E) = \begin{pmatrix} c_p(\mathcal{X}_E) + \lambda \nabla \phi(\pi)^\top - A_\star^\top \mu \\ A_\star \pi - (r, s_\star)^\top \end{pmatrix}.$$

Because $p \geq 2$, F is continuously differentiable in the neighborhood of a specific point $(\pi_{p,\lambda}, \mu_{p,\lambda}, (r_0, s_{0\star}), E_0)$ with $F(\pi_{p,\lambda}, \mu_{p,\lambda}, (r_0, s_{0\star}), E_0) = 0$. The matrix of the partial derivatives of F with respect to π and μ is given by

$$\nabla_{(\pi,\mu)} F(\pi_{p,\lambda}, \mu_{p,\lambda}, (r_0, s_{0\star}), E_0) = \begin{pmatrix} \lambda \nabla^2 \phi(\pi_{p,\lambda}) & -A_\star^\top \\ A_\star & 0 \end{pmatrix} \in \mathbb{R}^{(N^2+2N-1)\times(N^2+2N-1)}.$$

This matrix is non-singular because $\lambda > 0$, the Hessian $\nabla \phi(\pi_{p,\lambda})$ is positive definite (Section 2.1 in Klatt *et al.* (2020)) and the matrix $A_\star^\top$ has full rank. As a result, the implicit function theorem guarantees the existence of a continuously differentiable function that parameterizes the regularized optimal transport plan with projection in the neighborhood of $(r_0, s_{0\star}, E_0)$. The computation of the partial derivative form is directly followed by Theorem 2.3 and Example 2.6 in Klatt *et al.* (2020).

S1.4 Proof of Proposition 2

Since the regularized optimal transport distance is defined as

$$W_{p,\lambda}(r, s_\star; \mathcal{X}_E) = \langle c_p(\mathcal{X}_E), \pi_{p,\lambda}(r, s_\star; \mathcal{X}_E)\rangle^{1/p},$$

it follows from Lemma 1 that the map $(r, s_\star, E) \mapsto W_{p,\lambda}(r, s_\star, \mathcal{X}_E)$ is continuously differentiable on $\Delta_N \times (\Delta_N)_\star \times \mathbb{R}^{dk}$. Moreover, the matrix of partial derivatives with respect to (r, s_*) is given by

$$\nabla_{(r,s_\star)} W_{p,\lambda}(r, s_\star; \mathcal{X}_E) = \gamma^\top D A_\star^\top (A_\star D A_\star^\top)^{-1},$$

where γ is the gradient of function $\pi \mapsto \langle c_p(\mathcal{X}_E), \pi\rangle^{1/p}$ evaluated in the regularized transport plan $\pi_{p,\lambda}(r, s_\star; \mathcal{X}_E)$, which is formally defined by (3.16). Consequently, Theorem 3 implies that the map $(r, s_\star) \mapsto \mathrm{PW}_{p,\lambda}(r, s_\star)$ is directionally differentiable with derivative (3.15) in the sense of Gâteaux. To see this map is also a directionally derivative in the Hadamard sense, it is sufficient to show local Lipschitz continuity of this map (Proposition 3.5 of Shapiro (1990)). To this end, we fix a closed set $S_0 \subset \Delta_N \times (\Delta_N)_\star$. For any $(r, s_\star), (r', s'_\star) \in S_0$, we have

$$|\mathrm{PW}_{p,\lambda}(r, s_\star) - \mathrm{PW}_{p,\lambda}(r', s'_\star)| \leq \max_{E \in S_{d,k}} |W_{p,\lambda}(r, s_\star; \mathcal{X}_E) - W_{p,\lambda}(r', s'_\star; \mathcal{X}_E)|. \tag{S1.4}$$

Because the map $(r, s_\star, E) \mapsto W_{p,\lambda}(r, s_\star, \mathcal{X}_E)$ is continuously differentiable, there exists a constant $C > 0$ that does not depend on $(r, s_\star), (r', s'_\star)$ and

E so that

$$|W_{p,\lambda}(r, s_\star; \mathcal{X}_E) - W_{p,\lambda}(r', s'_\star; \mathcal{X}_E)| \leq C\|(r, s_\star) - (r', s'_\star)\|. \qquad \text{(S1.5)}$$

A combination of equations (S1.4) and (S1.5) leads to local Lipschitz continuity of the map $(r, s_\star) \mapsto \mathrm{PW}_{p,\lambda}(r, s_\star)$. This completes the proof.

S1.5 Proof of Theorem 4

The proof is a simple application of the delta method (Theorem 1), with the derivative of the regularized PRW distance (Proposition 2).

S1.6 Proof of Proposition 3

Under $r = s$, as shown in Proposition 1, the map $(r, s) \mapsto \mathrm{IW}_p^p(r, s)$ is directionally Hadamard differentiable. The proof is an application of Proposition 2 in Dümbgen (1993) in combination with the continuous mapping theorem. Under $r \neq s$, as discussed in the proof of Theorem 2, the map $(r, s) \mapsto \mathrm{IW}_p(r, s)$ is directionally Hadamard differentiable. The proof is a direct application of Proposition 2 in Dümbgen (1993)

S1.7 Proof of Proposition 4

As shown in Proposition 2, the map $(r, s) \mapsto \mathrm{PW}_{p,\lambda}(r, s)$ is directionally Hadamard differentiable. Then, the proof is a direct application of Proposition 2 in Dümbgen (1993) with this map.

S2 Additional Simulation Results

S2.1 Speed of convergence

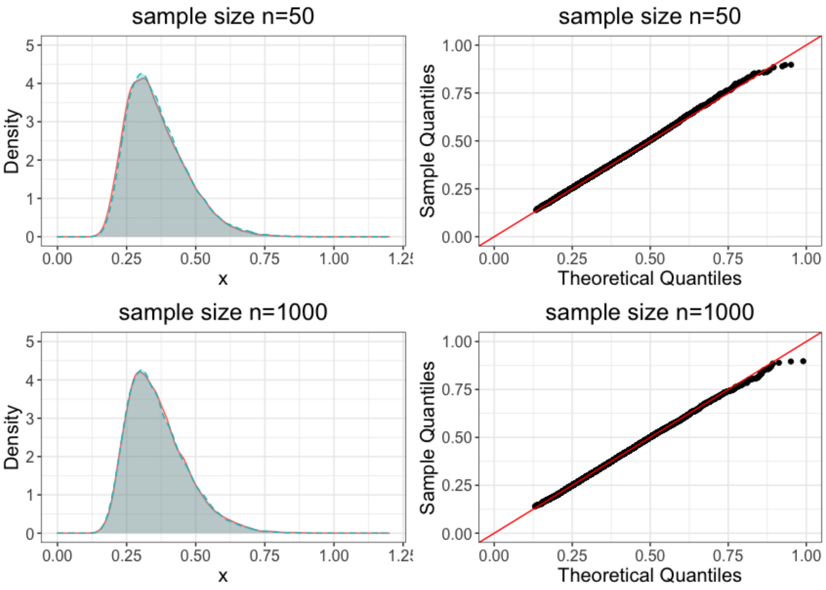
We illustrate our distributional limit results in Monte Carlo simulations. Specifically, we investigate the speed of convergence of the empirical IPRW distance ($p = 1$) and the empirical regularized PRW distance ($p = 2$) to their limit distributions (Theorems 2 and 4). All simulations were performed using R (Team *et al.* (2013)). The Wasserstein distances were calculated using the R package *transport* (Schuhmacher *et al.*, 2020), and the regularized transport distances were calculated using the R package *Barycenter* (Klatt, 2018).

Integral projection robust Wasserstein distance: We consider the finite ground space $\mathcal{X}$ to be an equidistant two-dimensional $L \times L$ grid on $[0,1] \times [0,1]$, with size $N = L^2$. We first set the grid size to $L = 7$ (i.e., $N = 49$).

For case $r = s$, we consider a probability distribution r on $\mathcal{X}$ as a realization of Dirichlet random variable $\mathrm{Dir}(\mathbf{1})$ with concentration parameter $\mathbf{1} = (1, ..., 1) \in \mathbb{R}^N$, and set $s = r$. Given such distributions $r, s \in \Delta_N$, we sample observations $X_1, ..., X_n \sim r$ and $Y_1, ..., Y_m \sim s$ i.i.d. with sample size $n = m \in \{25, 50, 100, 1000, 5000\}$ and compute $\sqrt{n/2}\mathrm{IW}_1(\hat{r}_n, \hat{s}_n)$ with one-dimensional projection and the uniform measure, which corresponds to the sliced Wasserstein distance. This process is repeated 20,000 times. Similarly, we consider the same setup for the case $r \neq s$, where we generate the second distribution, $s \sim \mathrm{Dir}(\mathbf{1})$, independently. We then compare the finite distributions with the theoretical limit distributions given by Theorem 2.

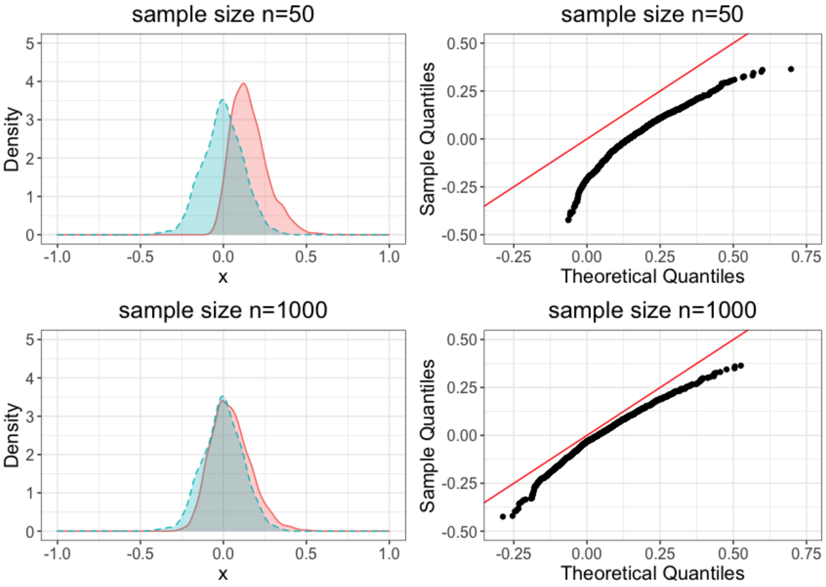
We demonstrate the results using kernel density estimators and Q-Q plots in Figure 1 and 2. The limit distributions are good approximations of finite sample distributions for a large sample size ($n = 1000$) in both $r = s$ and $r \neq s$. We also observe that, under $r = s$, the limit law approximates finite sample distribution quite well, even for a small sample size ($n = 50$). In Figure 3, we also show the speed of convergence with respect to the Kolmogorov–Smirnov distance (the maximum absolute difference between the distribution function of finite sample law and that of the limit law) for grid sizes $L = 3, 5, 7$. This shows that the Kolmogorov–Smirnov distances decrease as the sample size increases, and the size of ground space $N = L^2$ slows the speed of convergence marginally, especially for $r \neq s$.

Regularized projection robust Wasserstein distance: We consider the ground space $\mathcal{X}$ to be a form $\{1/M, 2/M, ..., M/M\} \times \{-0.001, 0.001\} \times$



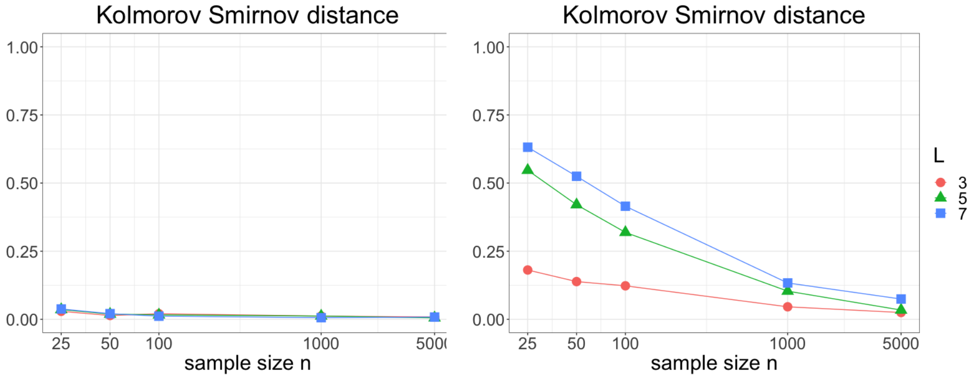
(a) $r = s$

Figure 1: **Comparison of finite sample distributions and the limit distribution of the empirical IPRW distance for the case $\mathbf{r} = \mathbf{s}$.** The first row shows finite sample density (dashed line) of the empirical IPRW distance for $n = 50$ on a regular grid of size $L = 7$ compared to its limit density (solid line). The densities are estimated by kernel density estimators with Gaussian kernel and Silverman's rule is used to select bandwidth. The corresponding Q-Q plot is presented on the right, where the solid line indicates perfect fit.The second row is the same setting as above, but $n = 1000$.



(a) $r \neq s$

Figure 2: **Comparison of finite sample distributions and the limit distribution of the empirical IPRW distance for the case $\mathbf{r} \neq \mathbf{s}$.** Same scenario as in Figure 1, but here the sampling distributions r and s are not equal.



(a) $r = s$

(b) $r \neq s$

Figure 3: **(A) Kolmogorov-Smirnov distances of the IPRW distance for the case $\mathbf{r = s}$.** The Kolmogorov-Smirnov distance between finite sample distributions of the empirical IPRW distance and its theoretical limit distribution for different sample size $n \in \{25, 50, 100, 1000, 5000\}$ and different grid sizes L. The axes are given on a logarithmic scale. **(B) Kolmogorov-Smirnov distances of the IPRW distance for the case $\mathbf{r \neq s}$.** Same scenario as in (A), but here the sampling distributions r and s are not equal.

$\{-0.001, 0.001\} \subset \mathbb{R}^3$ with grid size M and total size $N = 4M$. This ground space $\mathcal{X}$ is set to have a low-dimensional structure: two distributions on $\mathcal{X}$ differ mostly in the first coordinate, while the differences in the second and third coordinates are regarded as noise. For $M = 10$ (i.e., $N = 40$), we generated probability distributions r and s on $\mathcal{X}$ as realizations of independent Dirichlet random variables $\mathrm{Dir}(\mathbf{1})$. Given distributions $r \neq s$, we consider the same sampling scenarios as in the case of the IPRW distance and compute $\sqrt{n/2}\{\mathrm{PW}_{2,\lambda}(\hat{r}_n, \hat{s}_n) - \mathrm{PW}_{2,\lambda}(r, s)\}$ with one-dimensional projection and regularization parameter $\lambda = 1$. We repeat this process 20,000 times and compare finite distributions to its theoretical limit distribution given by Theorem 4.

Figure 4 shows the results demonstrated by kernel density estimators and Q-Q plots. The limit distributions are good approximations of finite sample distributions for both small and large sample sizes. Figure 5 shows the speed of convergence with respect to the Kolmogorov-Smirnov distance under grid sizes $M = 3, 7, 10$. We observe declining tendency of the Kolmogorov-Smirnov distances as the sample size increases.

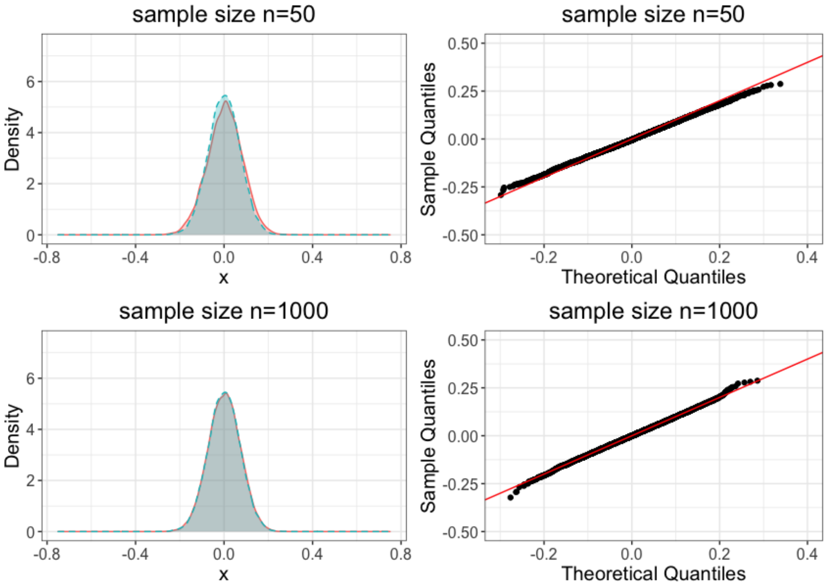


Figure 4: **Comparison of finite sample distributions and the limit distribution of the empirical regularized PRW distance for the case $\mathbf{r} \neq \mathbf{s}$.** The first row shows a finite sample density (dashed line) of the empirical regularized PRW distance for $n = 50$ on a ground space of grid size $M = 10$, which is compared to its limit density (solid line). The densities are estimated in the same way as Figure 1. The corresponding Q-Q plot is presented on the right, where the solid line indicates perfect fit. The second row is the same setting as above, but $n = 1000$.

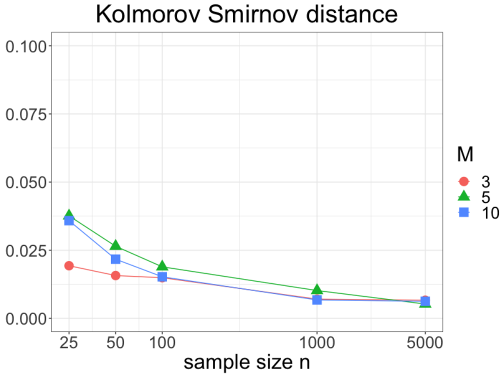
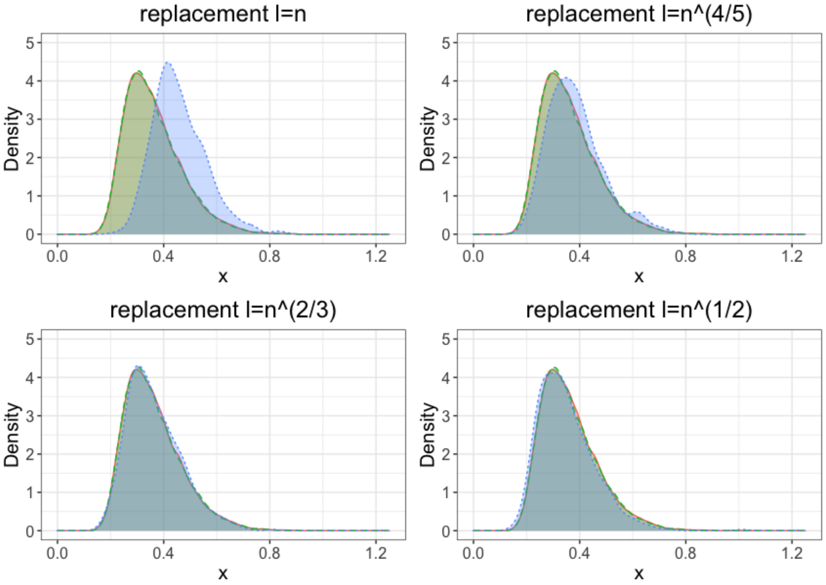


Figure 5: **Kolmogorov-Smirnov distances of the regularized PRW for $\mathbf{r} \neq \mathbf{s}$.** The Kolmogorov-Smirnov distance between finite sample distribution of the empirical regularized PRW distance and its theoretical limit distribution for different sample size $n \in \{25, 50, 100, 1000, 5000\}$ and different grid sizes M. The axes are given on a logarithmic scale.

S2.2 Simulation of bootstrap

We illustrate the acuuracy of approximation by the rescaled bootstrap (Proposition 3 and 4).

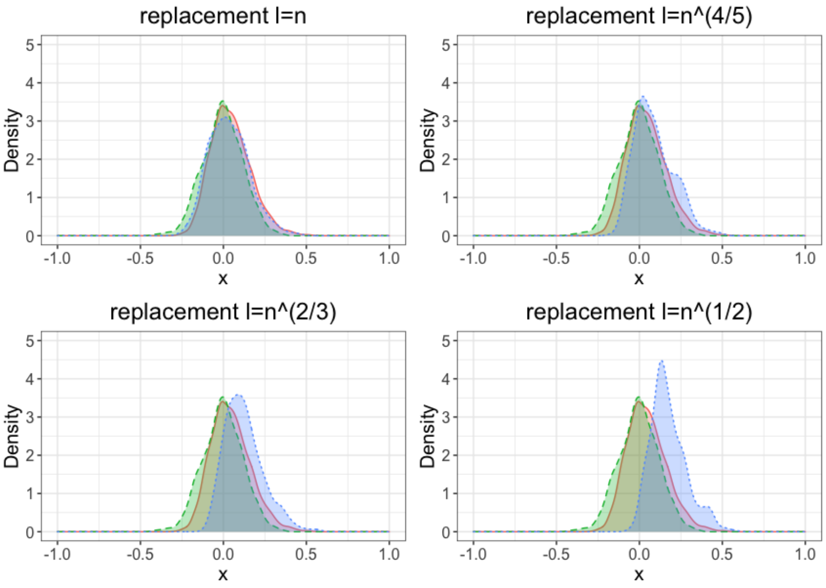
Integral projection robust Wasserstein distance: For a grid with $L = 7$, we generate $r \sim \mathrm{Dir}(\mathbf{1})$, set $s = r$, and sample $n = 1000$ observations according to probability distributions r, s. In addition, for fixed empirical distributions $\hat{r}_n, \hat{s}_n$, we generate $B = 500$ bootstrap replications of $\sqrt{\ell/2}\ \mathrm{IW}_1(\hat{r}^*_\ell, \hat{s}^*_\ell)$ by drawing independently with replacement $\ell \in \{n, n^{4/5}, n^{2/3}, n^{1/2}\}$ according to $\hat{r}_n$ and $\hat{s}_n$. Similarly, we consider the same setup in the case of $r \neq s$, where the second distribution, s, is generated independently from Dir($\mathbf{1}$). In the $r \neq s$ case, the form of bootstrap replications is $\sqrt{\ell/2}\{\mathrm{IW}_1(\hat{r}^*_\ell, \hat{s}^*_\ell) - \mathrm{IW}_1(\hat{r}_n, \hat{s}_n)\}$. The finite bootstrap sample distributions are then compared with their finite sample and theoretical limit distributions.



(a) $r = s$

Figure 6: **(A) Bootstrap for the empirical IPRW distance under $\mathbf{r = s}$.** Illustration of the rescaled plug-in bootstrap approximation ($n = 1000$) with replacement $\ell \in \{n, n^{4/5}, n^{2/3}, n^{1/2}\}$ and grid size $L = 7$. Finite bootstrap densities (dotted lines) are compared to their finite sample density (solid line) and limit density (dashed line). The densities are estimated in the same way with Figure 1.

The results are shown in Figure 6 and 7. We observe that, under $r = s$, finite bootstrap distributions with fewer replacements ($\ell = n^{4/5}, n^{2/3}, n^{1/2}$)



(a) $r \neq s$

Figure 7: **(B) Bootstrap for the empirical IPRW distance under $\mathbf{r \neq s}$.** Same scenario as in (A), but here the sampling distributions r and s are not equal.

are better approximations of the finite sample distribution than the naive bootstrap ($\ell = n$). This is consistent with the theoretical result in Section 4, which claims the naive bootstrap does not have consistency for the IPRW distance but resampling fewer observations leads to consistency. However, under $r \neq s$, the bootstrap approximations with fewer replacements are not good, and the naive bootstrap approximation is better. This good approximation by the naive bootstrap is possible due to the fact that the map $(r, s) \mapsto \mathrm{IW}_p(r, s)$ is only directionally Hadamard differentiable in general but (non-directionally) Hadamard differentiable at most points (r, s) with $r \neq s$. For instance, for ground size $N = 2$ (i.e., $\mathcal{X} = \{x_1, x_2\}$), the IPRW distance can be explicitly written as $\mathrm{IW}_p(r, s) = (\int_{S_{d,k}} \|E^\top(x_1 - x_2)\|^p d\mu(E))^{1/p}|r_1 - s_1|$. Therefore, in this case, the map $(r, s) \mapsto \mathrm{IW}_p(r, s)$ is Hadamard differentiable if $r \neq s$.

Regularized projection robust Wasserstein distance: For grid size $M = 10$, we generate distributions r and s as realizations of independent random variables from $\mathrm{Dir}(\mathbf{1})$ and sample $n = 1000$ observations according to probability distributions r, s. Additionally, for fixed empirical distributions $\hat{r}_n, \hat{s}_n$, we generate $B = 500$ bootstrap replications of $\sqrt{\ell/2}\{\mathrm{PW}_{2,\lambda}(\hat{r}_\ell^*, \hat{s}_\ell^*) - \mathrm{PW}_{2,\lambda}(\hat{r}_n, \hat{s}_n)\}$ with $\lambda = 1$ by drawing independently with replacement $\ell \in \{n, n^{4/5}, n^{2/3}, n^{1/2}\}$, according to $\hat{r}_n$ and $\hat{s}_n$. The finite bootstrap sample distributions are then compared with their finite sample and theoretical limit distributions.

The results are shown in Figure 8. The accuracy of the bootstrap approximation is not affected by replacement number ℓ in this case.

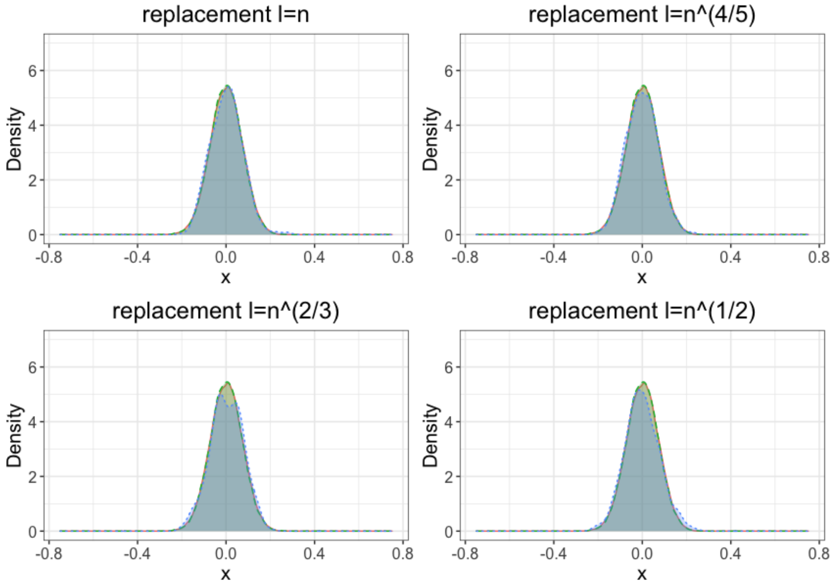


Figure 8: **Bootstrap for the regularized PRW distance under $\mathbf{r} \neq \mathbf{s}$.** Illustration of the rescaled plug-in bootstrap approximation ($n = 1000$) with the replacement $\ell \in \{n, n^{4/5}, n^{2/3}, n^{1/2}\}$ and grid size $M = 10$. Finite bootstrap densities (dotted lines) are compared with their finite sample density (solid line) and limit density (dashed line). The densities are estimated in the same way with Figure 1.

S3 Extension to Countable Spaces

In Section 3.2, we derived the limit distributions of empirical IPRW distances on finite spaces. In this section, we extend this result to countable spaces. Let $\mathcal{X} = \{x_1, x_2, ...\} \subset \mathbb{R}^d$ be a countable set and we assume that $\mathcal{X}$ is bounded. Let $\ell^1(\mathbb{R}^{\mathbb{N}}) = \{a \in \mathbb{R}^{\mathbb{N}} : \sum_{i \in \mathbb{N}} |a_i| < \infty\}$ be the set of absolutely summable sequences, and for $a \in \ell^1(\mathbb{R}^{\mathbb{N}})$, $\|a\|_{\ell^1} = \sum_{i \in \mathbb{N}} |a_i|$ be its norm. A probability measure on $\mathcal{X}$ is represented as an element in $\Delta_\infty = \{r \in \ell^1(\mathbb{R}^{\mathbb{N}}) : \sum_{i=1}^\infty r_i = 1, r_i > 0\}$. For $r, s \in \Delta_\infty$, the set of couplings is defined by

$$\Pi(r,s) = \left\{ \pi \in \ell^1(\mathbb{R}^{\mathbb{N}\times\mathbb{N}}) : A(\pi) = \begin{pmatrix} r \\ \\ s \end{pmatrix}, \pi \geq 0 \right\},$$

where $A : \ell^1(\mathbb{R}^{\mathbb{N}\times\mathbb{N}}) \to \ell^1(\mathbb{R}^{\mathbb{N}}) \times \ell^1(\mathbb{R}^{\mathbb{N}})$ is the marginalization operator, which is defined by

$$\pi \mapsto \begin{pmatrix} (\sum_{j=1}^\infty \pi_{ij})_{i=1}^\infty \\ \\ (\sum_{i=1}^\infty \pi_{ij})_{j=1}^\infty \end{pmatrix}.$$

The p-Wasserstein distance between two distributions $r, s \in \Delta_\infty$ on $\mathcal{X}$ is given by $W_p(r, s; \mathcal{X}) = \left\{\min_{\pi \in \Pi(r,s)} \langle c_p(\mathcal{X}), \pi \rangle\right\}^{1/p}$. Here, $c_p(\mathcal{X}) \in \ell^1(\mathbb{R}^{\mathbb{N}} \times \mathbb{N})$ denotes pair-wise transport costs such that $(c_p(\mathcal{X}))_{i,j} = \|x_i - x_j\|^p$, and the inner product $\langle c_p(\mathcal{X}), \pi \rangle = \sum_{i,j \in \mathbb{N}} \|x_i - x_j\|^p \pi_{ij}$ denotes the total costs associated to the transport plan $\pi \in \Pi(r, s)$. The Wasserstein distance between the projections of the distributions r, s in direction $E \in S_{d,k}$ is represented by $W_p(r, s; \mathcal{X}_E)$, where $\mathcal{X}_E = \{E^\top x_1, E^\top x_2, ...\} \subset \mathbb{R}^k$. Based on this, the p-IPRW distance between $r, s \in \Delta_\infty$ on $\mathcal{X}$ is represented as $\mathrm{IW}_p(r, s) = \left(\int_{S_{d,k}} W_p^p(r, s; \mathcal{X}_E) d\mu(E)\right)^{1/p}$ where μ is a given measure on $S_{d,k}$.

Let $r, s \in \Delta_\infty$ be probability measures on a bounded countable space $\mathcal{X}$. Let $\hat{r}_n, \hat{s}_m$ denote the empirical distributions generated by i.i.d. samples $X_1, ..., X_n \sim r$ and $Y_1, ..., Y_m \sim s$, respectively. We derive the limit distributions of $\sqrt{nm/(n+m)}\{\mathrm{IW}_p(\hat{r}_n, \hat{s}_m) - \mathrm{IW}_p(r, s)\}$ as $n, m \to \infty$. For description of our result, following Tameling *et al.* (2019), we define the following sets

$$\mathcal{S}^*(\mathcal{X}) = \{\lambda \in \ell^\infty(\mathbb{R}^{\mathbb{N}}) : \lambda_i - \lambda_j \leq \|x_i - x_j\|^p)$$

and

$$\begin{aligned}\mathcal{S}^*(r, s; \mathcal{X}) = \{&(\lambda, \mu) \in \ell^\infty(\mathbb{R}^{\mathbb{N}}) \times \ell^\infty(\mathbb{R}^{\mathbb{N}}) \\ &: \langle r, \lambda \rangle + \langle s, \mu \rangle = W_p^p(r, s; \mathcal{X}), \lambda_i + \mu_j \leq \|x_i - x_j\|^p\},\end{aligned}$$

where $\ell^\infty(\mathbb{R}^{\mathbb{N}}) = \{a \in \mathbb{R}^{\mathbb{N}} : \sup_{i \in \mathbb{N}} |a_i| < \infty\}$. Furthermore, for $r \in \Delta_\infty$, we define the following covariance structure

$$\Sigma(r) = \begin{cases} r_i(1 - r_i) & \text{if } i = j \\ -r_i r_j & \text{if } i \neq j. \end{cases}$$

Theorem S.1. *Let $\mathcal{X}$ be a bounded countable set, $r, s \in \Delta_\infty$ and $\hat{r}_n, \hat{s}_m$ be the empirical distributions generated by i.i.d. samples $X_1, ..., X_n \sim r$ and $Y_1, ..., Y_m \sim s$, respectively. Furthermore, let G and H be independent zero-mean Gaussian processes with covariance structures $\Sigma(r), \Sigma(s)$, respectively. Assume $\sum_{i=1}^\infty \sqrt{r}_i < \infty$ and $\sum_{i=1}^\infty \sqrt{s}_i < \infty$. Then, we have the followings:*

1. *If $r = s$, and $n \wedge m \to \infty$ and $m/(n+m) \to \delta \in (0,1)$, we have*

$$\left(\frac{nm}{n+m}\right)^{\frac{1}{2p}} \mathrm{IW}_p(\hat{r}_n, \hat{s}_m) \xrightarrow{d} \left(\int_{S_{d,k}} \sup_{\lambda \in \mathcal{S}^*(\mathcal{X}_E)} \langle G, \lambda \rangle d\mu(E)\right)^{1/p}.$$

2. *If $r \neq s$, and $n \wedge m \to \infty$ and $m/(n+m) \to \delta \in (0,1)$, we have*

$$\sqrt{\frac{nm}{n+m}}\{\mathrm{IW}_p(\hat{r}_n, \hat{s}_m) - \mathrm{IW}_p(r, s)\}$$
$$\xrightarrow{d} \frac{1}{p}\mathrm{IW}_p^{1-p}(r, s) \int_{S_{d,k}} \sup_{(\lambda,\mu) \in \mathcal{S}^*(r,s;\mathcal{X}_E)} \sqrt{\delta}\langle G, \lambda \rangle + \sqrt{1-\delta}\langle H, \mu \rangle d\mu(E).$$

To prove the claim of Theorem S.1, we use the delta method (Theorem 1 in Römisch (2004)). Therefore, we need to show directional Hadamard differentiability of $\mathrm{IW}_p^p(\cdot, \cdot)$ and weak convergence of $\sqrt{nm/(n+m)}\{(\hat{r}_n, \hat{s}_m) - (r, s)\}$. Directional Hadamard differentiability of IW_p^p is addressed in the following proposition.

Proposition S.1. *Assume that support $\mathcal{X}$ is a bounded countable set. Then the map IW_p^p from $(\Delta_\infty \times \Delta_\infty, \|\cdot\|_{\ell^1})$ to $\mathbb{R}$, $(r, s) \mapsto \mathrm{IW}_p^p(r, s)$ is directional Hadamard differentiable at all (r, s) tangentially to $\Delta_\infty \times \Delta_\infty$. The contin-*

gent cone on which the derivative is defined is given by

$$\mathcal{D}(r,s) = \mathcal{D}(r) \times \mathcal{D}(s)$$

with

$$\mathcal{D}(r) = \left\{ d \in \ell^1(\mathbb{R}^{\mathbb{N}}) \setminus \{0\} : \sum_{i=1}^{\infty} d_i = 0, d_i \in [-r_i, 1-r_i] \right\}.$$

Furthermore, the directional derivative is given as follows

$$(d_1, d_2) \mapsto \int_{S_{d,k}} \sup_{(\lambda,\mu)\in\mathcal{S}^*(r,s;\mathcal{X}_E)} -(\langle\lambda, d_1\rangle + \langle\mu, d_2\rangle) d\mu(E).$$

Proof of Proposition S.1. Consider a sequence $\{(h_{1\ell}, h_{2\ell})\} \subset \ell^1(\mathbb{R}^{\mathbb{N}})\times\ell^1(\mathbb{R}^{\mathbb{N}})$ with a limit $(h_1, h_2) \in \mathcal{D}(r,s)$ of the form $h_{1\ell} = t_\ell^{-1}(r_n - r), h_{2\ell} = t_\ell^{-1}(s_n - s)$ where $r_n, s_n \in \Delta_\infty$ and $t_\ell \searrow 0$. Following the definition of directional Hadamard derivative, we consider the following difference:

$$\begin{aligned} &\frac{\mathrm{IW}_p^p(r + t_\ell h_{1\ell}, s + t_\ell h_{2\ell}) - \mathrm{IW}_p^p(r,s)}{t_\ell}. \\ &= \int_{S_{d,k}} \frac{W_p^p(r + t_\ell h_{1\ell}, s + t_\ell h_{2\ell}; \mathcal{X}_E) - W_p^p(r,s;\mathcal{X}_E)}{t_\ell} d\mu(E), \end{aligned} \tag{S3.6}$$

and consider its limit. For each $E \in S_{d,k}$, Theorem A.3 in Tameling *et al.* (2019) implies

$$\frac{W_p^p(r + t_\ell h_{1\ell}, s + t_\ell h_{2\ell}; \mathcal{X}_E) - W_p^p(r,s;\mathcal{X}_E)}{t_\ell} \to \sup_{(\lambda,\mu)\in\mathcal{S}^*(r,s;\mathcal{X}_E)} -(\langle\lambda, h_1\rangle + \langle\mu, h_2\rangle),$$

as $\ell \to \infty$. The Lipschitz continuity of the Wasserstein distance on finite space (Theorem 4 of Sommerfeld and Munk (2018)) also holds even if the support $\mathcal{X}$ is a bounded countable set. Therefore, we have

$$\left|\frac{W_p^p(r+t_\ell h_{1\ell}, s+t_\ell h_{2\ell}; \mathcal{X}_E) - W_p^p(r,s;\mathcal{X}_E)}{t_\ell}\right| \leq \frac{p\mathrm{diam}(\mathcal{X}_E)^p \|t_\ell(h_{1\ell}, h_{2\ell})\|}{t_\ell}$$
$$\leq pk^p \mathrm{diam}(\mathcal{X})^p \|(h_{1\ell}, h_{2\ell})\|.$$

Because $\mathcal{X}$ is bounded and $\{h_{1\ell}\}$ and $\{h_{2\ell}\}$ are convergent sequences, the term $pk^p\mathrm{diam}(\mathcal{X})^p\|(h_{1\ell}, h_{2\ell})\|$ is bounded by a constant not depending on E and ℓ. Therefore, by taking $\ell \to \infty$ in (S3.6), we can apply the dominated convergence theorem, and the claim then holds. □

Proof of theorem S.1. Under the assumption that $\sum_{i=1}^\infty \sqrt{r}_i < \infty$ and $\sum_{i=1}^\infty \sqrt{s}_i < \infty$, Lemma 2.6 in Tameling *et al.* (2019) implies that $\sqrt{nm/(n+m)}\{(\hat{r}_n, \hat{s}_m) - (r,s)\} \xrightarrow{d} (\sqrt{\delta}G, \sqrt{1-\delta}H)$ with respect to the $\|\cdot\|_{\ell^1}$-norm, as $n/(n+m) \to \delta \in (0,1)$. Applying the delta method (Theorem 1 in Römisch (2004)) with $-G \overset{d}{=} G$ and $-H \overset{d}{=} H$, we have

$$\sqrt{\frac{nm}{n+m}}\{\mathrm{IW}_p^p(\hat{r}_n, \hat{s}_m) - \mathrm{IW}_p^p(r,s)\}$$
$$\xrightarrow{d} \int_{S_{d,k}} \sup_{(\lambda,\mu)\in\mathcal{S}^*(r,s;\mathcal{X}_E)} \sqrt{\delta}\langle G, \lambda\rangle + \sqrt{1-\delta}\langle H, \mu\rangle d\mu(E).$$

Under $r = s$, we obtain the desired result by applying the continuous mapping theorem for $f(x) = x^{1/p}$. Under $r \neq s$, we have the desired result by applying the delta method in combination with the chain rule for directional Hadamard differentiability (Proposition 3.6 in Shapiro (1990)). □

Bibliography

Bauschke, H. H., Combettes, P. L. *et al.* (2011) *Convex analysis and monotone operator theory in Hilbert spaces*, vol. 408, Springer.

Dümbgen, L. (1993) On nondifferentiable functions and the bootstrap, *Probability Theory and Related Fields*, **95**, 125–140.

Klatt, M. (2018) *Barycenter: Regularized Wasserstein Distances and Barycenters*, r package version 1.3.1.

Klatt, M., Tameling, C. and Munk, A. (2020) Empirical regularized optimal transport: Statistical theory and applications, *SIAM Journal on Mathematics of Data Science*, **2**, 419–443.

Römisch, W. (2004) Delta method, infinite dimensional, *Encyclopedia of Statistical Sciences*, **3**.

Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F. and Schmitzer, B. (2020) *transport: Computation of Optimal Transport Plans and Wasserstein Distances*, r package version 0.12-2.

Shapiro, A. (1990) On concepts of directional differentiability, *Journal of optimization theory and applications*, **66**, 477–487.

Sommerfeld, M. and Munk, A. (2018) Inference for empirical wasserstein distances on finite spaces, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**, 219–238.

Tameling, C., Sommerfeld, M. and Munk, A. (2019) Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications, *The Annals of Applied Probability*, **29**, 2744–2781.

Team, R. C. *et al.* (2013) R: A language and environment for statistical computing.