7 Supplemental material

In this supplement, we first prove that as $k/M \to 0$, $k, N \to \infty$ and for fixed s, the set $\mathcal{X}^*_{K,N}$ converges a.s. to the minimum ν-entropy set $\Omega_{1-\rho}$ containing a proportion of at least ρ of the mass of $f_0(x)$, where $\rho = \lim_{K,N\to\infty} K/N$ and $\nu = 1-\gamma/d$. We then derive asymptotic expressions for the bias and variance of the p-values $p_{bp}(.)$. Throughout this supplement, assume without loss of generality that $\{X_1, \ldots, X_N\} \in \mathcal{X}_N$ and $\{X_{N+1}, \ldots, X_T\} \in \mathcal{X}_M$.

7.1 Consistency

Denote the support of the density f_0 to be $\mathcal{S}$. Let $\mathcal{S}' \subset \mathcal{S}$ be any arbitrary subset of $\mathcal{S}$. Denote the collective behavior $k/M \to 0$. $k, N \to \infty$ by $\Delta \to 0$. Note that the distance $e_i(l)$ of a point $X_i \in \mathcal{X}_N$ to its l-th nearest neighbor in $\mathcal{X}_M$ is related to the bipartite l-nearest neighbor density estimate $\hat{f}_l(X_i) = \frac{l-1}{M c_d e_i^d(l)}$ (section 2.3, [12]) where c_d is the unit ball volume in d dimensions. From Theorem 3.1 and 3.2 in [12] it therefore immediately follows that

$$\lim_{\Delta\to 0} \mathbb{E}\left[\frac{1}{sN}\sum_{i=1}^{N} 1_{\{X_i\in\mathcal{S}'\}}(k/c_d M)^{\nu} d_{s,k}(X_i) - \int_{z\in\mathcal{S}'} f_0^{\nu}(z)\right]^2 = 0.$$

Because

$$\mathcal{X}^*_{K,N} = \underset{\mathcal{X}_{K,N}\in\mathcal{X}}{\operatorname{argmin}} L_{s,k}(\mathcal{X}_{K,N}, \mathcal{X}_M),$$

it follows that the set $\mathcal{X}^*_{K,N}$ converges to the minimum entropy set $\Omega_{1-\rho}$ containing a proportion of at least ρ of the mass of $f_0(x)$ where $\rho = \lim_{K,N\to\infty} K/N$.

7.2 Bias

Note that $\mathbb{E}[p_{orac}(X_0)] = p_{true}(X_0)$ and $\mathbb{V}[p_{orac}(X_0)] = p_{true}(X_0)(1-p_{true}(X_0))/N$. Let $\delta(X_i, X_0) = \delta_i = (f(X_i) - f(X_0))$. Also let $e(X) = \sum_{l=k-s+1}^{k} \hat{f}_l(X) - s f(X)$. We then have

$$\begin{aligned}
\mathbb{B}[p_{bp}(X_0)] &= \mathbb{E}[p_{bp}(X_0)] - p_{true}(X_0) = \mathbb{E}[p_{bp}(X_0) - p_{orac}(X_0)] \\
&= \mathbb{E}[1(d_{s,k}(X_1) \geq d_{s,k}(X_0))] - \mathbb{E}[1(f(X_1) \leq f(X_0))] \\
&= \mathbb{E}[1(e(X_1) - e(X_0) + \delta_1 \leq 0) - 1(\delta_1 \leq 0)].
\end{aligned}$$

This bias will be non-zero when $1(e(X_1) - e(X_0) + \delta_1 \leq 0) \neq 1(\delta_1 \leq 0)$. First we investigate this condition when $\delta_1 > 0$. In this case, for $1(e(X_1) - e(X_0) + \delta_1 \leq 0) \neq 1(\delta_1 \leq 0)$, we need $-e(X_1) + e(X_0) \geq \delta_1$. Likewise, when $\delta_1 \leq 0$, $1(e(X_1) - e(X_0) + \delta_1 \leq 0) \neq 1(\delta_1 \leq 0)$ occurs when $e(X_1) - e(X_0) > |\delta_1|$.

From the theory developed in Appendix C in [12], $|e(X)| = O(k/M)^{1/d} + O(1/\sqrt{k})$ with probability greater than $1 - o(1/M)$. This implies that

$$\begin{aligned}
\mathbb{B}[p_{bp}(X_0)] &= \mathbb{E}[1(e(X_1) - e(X_0) + \delta_1 \leq 0) - 1(\delta_1 \leq 0)] \\
&= Pr\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} + o(1/M).
\end{aligned} \tag{4}$$

We first analyze the case where f_0 is monotonic. By the continuity of f_0, we then have $||X_1 - X_0||^d = O(\delta_1)$. Because we assume the density f_0 is bounded above by some constant C on its support, we have

$$\begin{aligned}
Pr\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} &= Pr(||X_1 - X_0||^d = O(k/M)^{1/d} + O(1/\sqrt{k})) \\
&= O(k/M)^{1/d} + O(1/\sqrt{k}).
\end{aligned} \tag{5}$$

We now extend this analysis to the general case where f_0 is assumed to have a finite number of modes. Let $S_{X_0}(\delta) = \{X \in \mathcal{S} : |f(X) - f(X_0)| < \delta\}$. By the continuity of f_0, the volume

$V_{X_0}(\delta) = \int_{S_{X_0}(\delta)} dx = O(\delta)$. We then have

$$\begin{aligned}
\mathbb{B}[p_{bp}(X_0)] &= \mathbb{E}[1(e(X_1) - e(X_0) + \delta_1 \leq 0) - 1(\delta_1 \leq 0)] \\
&= Pr\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} + o(1/M) \\
&= Pr\{X_1 \in S_{X_0}(O(k/M)^{1/d} + O(1/\sqrt{k}))\} + o(1/M) \\
&= O(V_{X_0}(O(k/M)^{1/d} + O(1/\sqrt{k})) \\
&= O((k/M)^{1/d} + 1/\sqrt{k}).
\end{aligned}$$

7.3 Variance

Define $b_i = 1(e(X_i) - e(X_0) + \delta_i \leq 0) - 1(\delta_i \leq 0)$. We can compute the variance in a similar manner to the bias as follows

$$\begin{aligned}
\mathbb{V}[p_{bp}(X_0)] &= \frac{1}{N}\mathbb{V}[1(d_{s,k}(X_1) \geq d_{s,k}(X_0))] \\
&\quad + \frac{N-1}{N} Cov[1(d_{s,k}(X_1) \geq d_{s,k}(X_0)), 1(d_{s,k}(X_2) \geq d_{s,k}(X_0))] \\
&= \frac{1}{N}\mathbb{V}[1(e(X_1) - e(X_0) + \delta_1 \leq 0)] + \frac{N-1}{N} Cov[b_1, b_2] \\
&= O(1/N) + \mathbb{E}[b_1 b_2] - (\mathbb{E}[b_1]\mathbb{E}[b_2]) \\
&= O(1/N) + Pr\{\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} \cap \{|\delta_2| = O(k/M)^{1/d} + O(1/\sqrt{k})\}\} \\
&\quad + Pr\{|\delta_1| = O(k/M)^{1/d} + O(1/\sqrt{k})\} Pr\{|\delta_2| = O(k/M)^{1/d} + O(1/\sqrt{k})\} + o(1/M) \\
&= O(1/N + (k/M)^{2/d} + 1/\sqrt{k}).
\end{aligned}$$