# Hypothesis Tests with Functional Data for Surface Quality Change Detection in Surface Finishing Processes

Shilan Jin[a], Rui Tuo[a], Akash Tiwari[a], Satish Bukkapatnam[a], Chantel Aracne-Ruddle[b], Ariel Lighty[b], Haley Hamza[b], Yu Ding[a]†

[a]Department of Industrial & System Engineering,
Texas A&M University, College Station, Texas, USA
[b]Lawrence Livermore National Lab, Livermore, California, USA
†Correspondence author, yuding@tamu.edu

## 1 Appendix: Type II Error.

Recall that after discretizing the continuum of comparisons, taking $H_1^{\mu_{up}}$ in Equation (14) as an instance, we obtain $|D_u|$ tests that constitute a multiple testing, where $|D_u|$ denotes the cardinality of the set $D_u$. Westfall and Young (1993) (in Chapter 6) state "*in multiple testing situations, power is not so easily defined*." They list four definitions of the power of multiple testing. Using the notation of our hypothesis statement in Equation (14), the four definitions of power listed by Westfall and Young (1993) are following:

1. the probability of correctly rejecting at least one $H_0^{\mu_{up}}(s)$, for $s \in D_u$,
2. the probability of correctly rejecting all $H_0^{\mu_{up}}(s)$, for $s \in D_u$,
3. the probability of correctly rejecting exactly one $H_0^{\mu_{up}}(s)$, for $s \in D_u$, and
4. for a specific $s \in D_u$, the probability of correctly rejecting $H_0^{\mu_{up}}(s)$.

Westfall and Young (1993) analyzed the power in Definition 1 and confirmed the permutation test leads to a superior performance. As we follow the same permutation test strategy as suggested by Westfall and Young (1993), we will not repeat the analysis of power under Definition 1. Instead, we analyze the power in Definition 2, also because Definition 2 suits our hypothesis statement better. We do so using a simulation experiment; otherwise, the ground truth is unknown and the probability of *correctly* rejecting cannot be discerned. The power of a test is directly related to its type-II error, i.e., power = $1-$ type-II error.

We simulate two groups of $N$ functions to estimate the type II error of the mean test for upper tail and lower tail, while controlling the type I error to be under the nominal level, i.e., $\alpha = 0.03$. The two groups of functions to be tested are simulated from a Gaussian process, with one group of functions digressing from the other group by a small perturbation. To quantify the small perturbation between two groups of functions, we use a $L^2$-distance percentage defined as:

$$L^2\% = \frac{\|\mu_1 - \mu_2\|_{L^2}}{\|\mu_1\|_{L^2}} \times 100\%, \tag{1}$$

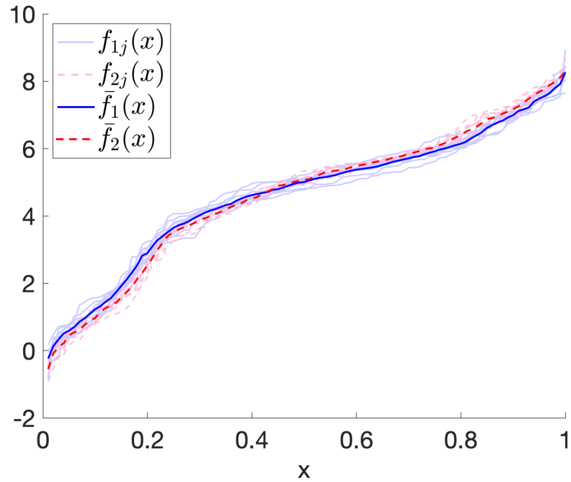


Figure 1: The two groups of nine functions, $f_{1j}(x)$ and $f_{2j}(x)$, simulated from one run, and the mean functions, $\bar{f}_1(x)$ and $\bar{f}_2(x)$.

where $\mu_i$ is the mean function of the functions of group $i$.

To generate the two groups of $N$ functions, we randomly sample one set of input points, $x \in [0, 1]$, as the pointwise test requires that two groups of functions have to be evaluated at the same input points. The two groups of functions are generated from the model described as: $f_{ij}(x) = z(x) + \epsilon_{ij}$, $i \in \{1, 2\}$, $j \in \{1, \cdots, N\}$; $z(x) \sim GP(\mathbf{0}, k(x, x'))$; $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ and $\sigma_\epsilon = 0.5$. The covariance function $k(x, x')$ is a squared exponential kernel function with the form: $k(x, x') = \sigma_f^2 \exp(-0.5[(x - x')/\theta]^2)$. We set $\sigma_f = 5$ and $\theta = 0.2$. To make the second group of functions that deviates from the first group, we add a perturbation $\delta(x)$ to $f_{2j}(x)$, $j \in \{1, \cdots, N\}$. The perturbation function $\delta(x)$ is created as:

$$\delta(x) = \begin{cases} -\frac{1}{3}\sin\left(\pi\left(\frac{x-0.2}{0.8-0.2}\right)\right), & x \leq 0.25, \\ 0, & 0.25 < x < 0.75, \\ \frac{1}{3}\sin\left(\pi\left(\frac{x-0.2}{0.8-0.2}\right)\right), & x \geq 0.75. \end{cases} \tag{2}$$

Thus, $f_{1j}(x) = z(x) + \epsilon_{1j}$ and $f_{2j}(x) = z(x) + \delta(x) + \epsilon_{2j}$, $j = 1, \cdots, N$. We generate $N = 6, 9, 12, 15$ functions for each of the two groups by sampling $\epsilon_{ij}$. Run our proposed hypothesis test for the mean tails and repeat the process for $1,000$ runs. The functions $f_{1j}(x)$ and $f_{2j}(x)$ generated for one run and their mean functions $\bar{f}_1(x)$ and $\bar{f}_2(x)$ are shown in Figure 1.

The average $L^2$ distance percentage over between two groups of functions over $1,000$ runs is $3.52\%$. The type II error ($\beta$) of the hypothesis test on the simulation data, with the type I error controlled under $0.03$, varies with the number of curves, $N$. Table 1 shows the estimated out-of-control average run length (ARL), that is defined as $1/(1 - \beta)$ and the average $L^2$ distance percentage corresponding to each $N$.

The type II error decreases as the number of curves of each group increases and so the average run length decreases, while the distance between two groups of curves stays more or less the same. That is consistent with the commonsense of a large sample size increasing the power of a hypothesis test. The out-of-control ARL results say that the HT-based method reacts to signal the small shift from group 1 to group 2 with a one to two sample delay on average. The small ARL values indicate effective detection.

In order to further reduce the type-II error or the out-of-control ARL, one needs to increase the sample size. However, in practice it is rather cumbersome and, sometimes, infeasible to take many samples. As in our polishing experiment, we are only able to image nine to $15$ locations on the spherical surface of

Table 1: Estimated type II errors for the simulated $N$ curves, with type I error under control.

| Number of curves, $N$ | Average $L^2$ distance percentage | Out-of-control average run length | |
|---|---|---|---|
| | | Mean test for upper tail | Mean test for lower tail |
| 6 | 3.61% | 1.29 | 1.62 |
| 9 | 3.54% | 1.18 | 1.37 |
| 12 | 3.49% | 1.18 | 1.30 |
| 15 | 3.43% | 1.10 | 1.21 |

peppercorn-sized bead. The sample sizes in Table 1 are included with this practical constraint in mind.

# References

Westfall, P. H. and S. S. Young (1993). Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. New York, NY: John Wiley & Sons.