Hypothesis tests with functional data for surface quality change detection in surface finishing processes

Shilan Jin\textsuperscript{a}, Rui Tuo\textsuperscript{a}, Akash Tiwari\textsuperscript{a}, Satish Bukkapatnam\textsuperscript{a},\textsuperscript{b}, Chantel Aracne-Ruddle\textsuperscript{b}, Ariel Lighty\textsuperscript{b}, Haley Hamza\textsuperscript{b}, and Yu Ding\textsuperscript{a}\textsuperscript{c}

\textsuperscript{a}Department of Industrial & System Engineering, Texas A&M University, College Station, TX, USA; \textsuperscript{b}Lawrence Livermore National Lab, Livermore, CA, USA

\textbf{ABSTRACT}

This work is concerned with providing a principled decision process for stopping or tool-changing in a surface finishing process. The decision process is supposed to work for products of non-flat geometry. The solution is based on conducting hypothesis testing on the bearing area curves from two consecutive stages of a surface finishing process. In each stage, the bearing area curves, which are in fact the nonparametric quantile curves representing the surface roughness, are extracted from surface profile measurements at a number of sampling locations on the surface of the products. The hypothesis test of these curves informs the decision makers whether there is a change in surface quality induced by the current finishing action. When such change is detected, the current action is deemed effective and should thus continue, while when no change is detected, the effectiveness of the current action is then called into question, signaling possibly some change in the course of action. Application of the hypothesis testing-based decision procedure to both spherical and flat surfaces demonstrates the effectiveness and benefit of the proposed method and confirms its geometry-agnostic nature.

\section{1. Introduction}

In precision manufacturing, polishing is an inevitable post-processing step towards ensuring a nano-scale surface finish for manufactured objects when the surface roughness requirement goes beyond the capability of the manufacturing operations prior to polishing. DeJule (1997) and Frazier (2014), among others, discussed the necessity and impact of polishing. A polishing process employs a polishing tool (an abrasive-embedded pad) that performs a repetitive rubbing action to remove the asperities on an object’s surface in a small amount at each action. An effective and efficient polishing process requires changing the polishing tool from time to time, as the coarse tools cannot deliver a fine finish, but polishing with fine tools would progress too slowly. The tool also steadily degrades during the process due to loading, glazing and other issues (Rao et al., 2015). The decisions of when to change the polishing tool or to stop polishing altogether are crucial, demanding sound and consistent guidance (Bukkapatnam et al., 2018; Jin et al., 2020).

In our research, we polish diamond-coated silicon balls of radius 1.688 mm to make its surface roughness to the order of \(~\text{10 nm in } \text{Sa (ISO 4287, 1997).}\) The balls are used in the physical experiments for proving the feasibility of inertial confinement fusion (Biener et al., 2009). During the experiments, the initial ultraviolet laser surrounding the balls is converted into soft X-rays. Bursts of the soft X-rays drive the compression of the materials in the balls to conditions similar to those found deep in the sun. The success of these experiments highly depends on the quality of the ball and its surface smoothness. Figure 1 presents a picture of the balls for a visual comparison of the surface before and after being polished. The left ball is unpolished and the right one is polished to about 12 nm in Sa. The picture was taken by a Leica DM4000 M LED with transmitted light. The polished ball reflects the ring illumination as a white circle on its surface.

The polishing process is carried out over \(T\) stages, of which each stage denotes a period of time when a chosen polishing tool is applied to the surface of the polished object under the chosen conditions (like the force, rotary speed of the ball mandrel, etc.). After each stage, the polishing action is paused and the surface is inspected by a microscopic measurement or imaging device, such as an optical profilometer, before the polishing action continues for the next stage. This polishing-inspection iteration repeats for every stage. What is measured after each stage affects the polishing decisions outlined earlier.

In our specific ball polishing process, during an inspection, the profilometer measurements are taken at \(M\) locations over the surface of the polished object (the white dots in Figure 2, top-left panel). For each location, the profilometer returns the surface morphology covering a small area of 229.76 \(\times\) 172.32 \(\mu m^2\), discretized into 640 \(\times\) 480 pixels over
the \( X \)-\( Y \) plane. On each pixel, a height value, denoted by \( z \) and in the unit of \( \mu m \), is registered by the profilometer. After some data preprocessing, which is primarily to calibrate the pixel height relative to its local surface baseline, the calibrated pixel heights are arranged into a matrix, with its row and column corresponding to the \( X \) and \( Y \) coordinates and its entry corresponding to the \( z \) value. In engineering practice, this matrix representation is further transformed into a one-dimensional curve, with the pixel heights sorted from the highest peak to the deepest valley; see Figure 2, bottom-left panel. Such a curve is known as the Bearing Area Curve (BAC) in manufacturing (Stewart, 2000) or a quantile curve in statistics. With such a data arrangement, the surface quality at any stage is represented by a group of BACs (quantile curves), each of which corresponds to one of the white spots.

Comparing the statistical characteristics of these nonparametric BACs of two successive stages could inform us whether there is a surface quality change, caused by the polishing action between the two inspections. When the action does make a difference to an object’s surface, it most likely results in the removal of peaks and valleys, that are reflected by the upper and lower tail of the BACs, respectively. As a result, one expects to see the tails of the mean curves to be flattened (or more precisely, upper tails lowered and lower tails raised) and/or the variances to be reduced.

Figure 3 presents two examples illustrating the observations. Figure 3(a) is at a stage where the polishing causes an obvious change in the mean curves but not so in the variance curves, whereas Figure 3(b) is at a stage where the polishing improves both the mean and variance curves but the improvement in variance is more pronounced. The detection

---

Figure 1. The silicon balls before and after being polished.

Figure 2. Top-left: the ball surface is inspected at the sample locations (white dots). Top-right and bottom-right: each location is measured by an optical profilometer which returns a matrix of pixel heights. Bottom-left: sorting the pixel heights yields a bearing area curve or the pixel-height quantile curve.
of surface change may directly inform about polishing decisions. When a polishing action leads to detectable surface changes, the implication is that it is effective, and consequently, the same action should be continued. When a polishing action does not lead to detectable surface changes, it signals the need to clean or change the current tool (to a finer scale), or if the current tool is already at the finest scale, the time to stop polishing.

Motivated by this observation, we formulate the polishing decision problem to be based on the change detection in nonparametric functional curves. Let \( \mu_t(s) \) and \( \sigma_t^2(s) \) be the mean and variance curve of the BACs at Stage \( t \), respectively, where \( s \) is the quantile variable. Our specific formulation consists of testing the following conditions:

\[
\begin{align*}
\mu_t(s) &< \mu_{t-1}(s) \quad \text{for any } s \in [0, \tau], \quad \text{or} \\
\mu_t(s) &> \mu_{t-1}(s) \quad \text{for any } s \in [1-\tau, 1], \quad \text{or} \\
\sigma_t^2(s) &< \sigma_{t-1}^2(s) \quad \text{for at least 50\% of } s \in [0, 1],
\end{align*}
\]

where \( \tau \) is a quantile cut-off for identifying the upper and lower tails. For simplicity, we use the same \( \tau \) for both tails, although it is not difficult to use two different quantile cut-offs. In the variance test, we test the variance reduction for at least 50\% of the values because there is no guarantee that the variance is always uniformly reduced (although the right panel in Figure 3(b) shows so). Using 50\% in the test, we deem the condition true if there are more instances of reduction than otherwise.

Apparently, hypothesis testing methods on nonparametric functional curves benefit directly our engineering decision problem; for this reason, we will provide a detailed review on nonparametric curves testing in Section 2. In our effort, we make a particular use of the pointwise testing method developed by Cox and Lee (2008). The original method in Cox and Lee (2008) tests the equality of two-sample mean curves, assuming the two samples have a common covariance structure. We extend their original method to the testing of both mean and variance curves for certain sub-domains of the curves.

The merit of our research can be summarized as follows. The proposed statistical testing does not require the sample location information and can thus be applicable to manufactured object of non-flat surface for which the location information is difficult to register. This applicability to non-flat surface, specifically a spherical surface in this study, is a major advantage in practice and substantially broadens the scope of applications. The statistical testing method is able to signal subtle and detailed changes in surface roughness and appears to be a well-suited tool for the much needed polishing decisions. Because the resulting method is based on nonparametric curves, it relies on few assumptions and appears robust and easy to use. The statistical testing method, when applied to the polishing data, confirms the phenomenon of over-polishing, i.e., when the same polishing action is used for an excessively long time, it harms the surface quality rather than improving it. This finding reinforces the importance of timely decisions on tool changing or stopping. Making timely decisions leads furthermore to significant saving in time, materials and energy in a polishing process. Although we use polishing to motivate the study, our proposed procedure is applicable to other surface finishing actions. In the latter case study, we present the analysis using both polishing and lapping data.

The rest of this article is organized as follows. In Section 2, we review both the current practice of polishing decision making and the curve testing literature. Section 3 presents the proposed hypothesis testing-based method for detecting surface quality change. Section 4 applies the hypothesis testing-based method to three polishing/lapping experiments and demonstrates the impact it makes in terms of enhanced process decisions. Section 5 concludes this work.

2. Literature review

In this section, we review the literature in terms of both decision making in polishing processes and testing the difference between nonparametric functional curves.

2.1. Decision making in polishing processes

Our focus here is about the methods based on direct surface roughness measurements, also known as product measurements-based methods. Quantification of manufactured surface quality remains an unresolved issue despite the efforts during
the past decades (Jiang et al., 2007a). For a long time and even nowadays, the benchmark method used in practice has been based on $Sa$ (or $Ra$ if it is concerned with a one-dimensional line feature), which is an average roughness metric defined as the mean absolute deviation of a surface’s roughness. Recall the pixel height matrix of $z$ values over the $X$-$Y$ plane at a give location, as illustrated in the bottom-right plot in Figure 2. The $Sa$ for a location is calculated as

$$\text{Sa} = \frac{1}{A} \int_{X,Y} |z(X, Y) - \bar{z}|dXdY,$$  

(2)

where $\bar{z}$ is the sample mean of $z(X, Y)s$ at the location and $A$ is the size of the area over which $z$ values are obtained. One $Sa$ is calculated for each location, i.e., each of the white dots in the top-left panel in Figure 2. When the whole surface has $M$ sample locations, i.e., $M$ white dots, there are $M$ distinct $Sa$ values. Practitioners often use the median of the $M$ measurements of $Sa$ values, denoted by $\overline{\text{Sa}}$, to benchmark the roughness for the current stage of operations. The popularity of median $Sa$ in practice is due to its easiness to compute/to use and its straightforward interpretability. However, recent works (Wang et al., 2014; Liu et al., 2017; Bukkapatnam et al., 2018; Jin et al., 2020) showed peoples’ awareness of a number of critical limitations of using average metric such as $Sa$ or median $Sa$ for the finishing process of precision manufacturing. Jin et al. (2020) provided specific examples where a set of similar media $Sa$ values correspond to surfaces with rather different roughness features.

Several alternative quantifiers of product surfaces have been pursued in the last five decades (Jiang et al., 2007b). These quantifiers include frequency domain, multi-scale and topological quantifiers (Jiang, 2012). Whitehouse (1982, 2001) argues that most of these quantifiers constitute a proliferation of parameters—very few of them are meaningful in terms of informing the functionality of surface characterization. Ganti and Bhushan (1995) and Jiang et al. (2007a) explain the importance of surface characterization and the lack of studies in this aspect. Even fewer quantifiers capture how a manufacturing process modifies the surface (Rao et al., 2015; Iquebal et al., 2017; Bukkapatnam et al., 2018).

One of the most recent additions in the category of product measurements-based methods is by Jin et al. (2020). They proposed a decision making criterion for polishing by modeling the surface roughness through a Gaussian Process (GP). The polished object surface is treated like a landscape, so that its analogy to spatial statistics is invoked, explaining why the GP model is used. They demonstrated that the scale parameter in the resulting GP model is sensitive to changes in the surface roughness and consequently devised a decision procedure based on that. While Jin et al. (2020) took an important step forward in introducing sophisticated statistical modeling to decision making in polishing, there are still two limitations in their method.

The first is that the use of a single scale parameter in the GP model as the representation of the surface roughness still compresses the detailed surface quality information and may lead to information loss in subsequent decision processes. The second limitation is that the applicability of the resulting GP-based method is limited to flat (or nearly flat) surfaces, due to the use of a squared-exponential covariance structure in the GP model. When the surface has a strong curvature, like the ball-surface polishing here, or in some other cases where complicated geometric features are involved, either the covariance function in the GP model must have a fundamental redesign (to account for the effect of geodesic distances) or a new geometry-independent decision process needs to be designed.

Accommodating the aforementioned two requirements, we propose a new non-flat surface-applicable method, which is based on hypothesis testing of nonparametric quantile curves (the BACs) associated with two consecutive surface finishing stages.

### 2.2. Hypothesis tests with nonparametric functional data

Hypothesis tests of nonparametric functions entail the mean function and covariance structure tests. In terms of the mean function test, Hall and Hart (1990) studied a test statistic produced by the difference of the estimations of the two functions, and proposed to use bootstrap to estimate the distribution of the test statistic. King et al. (1991) modified the test statistic of Hall and Hart (1990), so that an asymptotic distribution can be attained under the Gaussian assumption. Kulasekera (1995), Munk and Dette (1998) and Neumeyer and Dette (2003) investigated their respective test statistics. Chapters 5 and 9 in Zhang (2013) tested the mean functions of the samples of two curves through testing the $L_2$-difference between the two mean functions.

In terms of variance tests, the common technique is to use Karhunen–Loève expansion to approximate the continuous covariance operators with a finite set of Functional Principal Components (FPCs), and test the equality of each component between two samples. Benko et al. (2009) tested the equality tests of both mean functions and covariance structures. They constructed the test statistics that have been the distance of the two-sample FPCs and used bootstrap to estimate the distributions of the test statistics. Some other works have been devoted to the testing of the covariance structures. For instance, Panaretos et al. (2010) constructed a test statistic to test the FPCs assuming that the functional data follows GPs. Fremdt et al. (2013) tested the same type of problem but relaxed the GP assumption. Chapter 10 of Zhang (2013) discussed the test of two covariance functions by defining an $L_2$-based test statistic, measuring the difference between two covariance matrices.

The above-referenced works produce outcomes for a so-called global test, meaning that they give a binary answer concerning whether the two sets of functional curves are the same or not. However, they did not identify where the differences may lie, nor do these methods work for a subset of the input domain. Recall that the hypothesis tests we envision for the polishing process need to test for the tail portions or a subdomain (at least 50%), rendering these methods not directly applicable to our engineering problem.

To the best of our knowledge, there are two studies that can do a local test, i.e., the test can be done to a
subdomain, as the nature of our problem requires. One is Cox and Lee (2008) and the other is Prakash et al. (2022). Cox and Lee (2008) studied the testing problem when there are curve replicates and the input locations where the curves that are sampled are the same between the two groups of curves. Prakash et al. (2022) took advantage of the Bayesian posterior covariance to build a confidence band for the mean function difference. Their method is applicable to the circumstance where there is no curve replicate (a single curve in each test group) and the input locations are different.

Our problem setting matches with that of Cox and Lee (2008) to a greater extent. Thus, we choose to follow Cox and Lee (2008). However, Cox and Lee (2008) only presented a mean function test over the whole domain. To solve our problem, we need to introduce a mean function test for the tail portions and a variance test. Furthermore, we need to combine all the tests for devising a unified decision rule for making polishing decisions.

3. Tests and the detection rule

We discuss how we expand the hypothesis test proposed by Cox and Lee (2008), so that the testing method can be applied to the inequality test of two-sample mean functions and that of two-sample variance functions. The new test provides the basis for designing a rule for detecting the change in surface quality.

The basic idea of Cox and Lee (2008) is to approximate the curve test with multiple pointwise tests on the curve, and considering the multiple tests as a family. As the family of tests are highly correlated when the points are close to each other, Cox and Lee (2008) use the permutation method of Westfall and Young (1993) to control the familywise error rate. We extend the method of Cox and Lee (2008) and establish test statistics corresponding to the hypothesis tests specifically required for the polishing application.

3.1. Hypothesis tests

Two groups of curves are represented with \( z_{0j}(x) \), where \( t = 1, 2 \) is the group index and \( j = 1, \ldots, J \) is the curve index. Without loss of generality, let \( x \in [0, 1] \), meaning the input value is normalized to the range of \([0, 1]\), We want to test for the existence of an inequality between \( \mu_1(x) \) and \( \mu_2(x) \), that is the population mean of two groups of curves, and for the existence of an inequality between \( \sigma^2_1(x) \) and \( \sigma^2_2(x) \), that is the population variance of two groups of curves, for a certain subdomain of the curves. Denote by \( \varphi(x) \) either \( \mu(x) \) or \( \sigma^2(x) \). The subdomain is denoted by \( A \subset [0, 1] \). There are multiple options for this subdomain \( A \); for instance, it could be all \( x \in [0, 1] \), at least 50% of \( x \in [0, 1] \), or at least one \( x \in [0, 1] \). The general form of the hypothesis test of functional inequality can be expressed as

\[
H^o_0 : \varphi_1(x) = \varphi_2(x) \text{ for all } x \in [0, 1] \\
\text{against } H^o_1 : \varphi_1(x) > \varphi_2(x) \text{ for } x \in A.
\]

A few specific hypotheses are given in the following:

1. Test on the mean function, regardless of the relationship between \( \sigma^2_1(x) \) and \( \sigma^2_2(x) \):

\[
H^o_0 : \mu_1(x) = \mu_2(x) \text{ for all } x \in [0, 1] \\
\text{against } H^o_1 : \mu_1(x) > \mu_2(x) \text{ for all } x \in [0, 1].
\]

In this example, \( \varphi(x) = \mu(x) \) and \( A = \{ x : \text{ all } x \in [0, 1] \} \).

2. Test on the variance function, regardless of the relationship between \( \mu_1(x) \) and \( \mu_2(x) \):

\[
H^o_0 : \sigma^2_1(x) = \sigma^2_2(x) \text{ for all } x \in [0, 1] \\
\text{against } H^o_1 : \sigma^2_1(x) > \sigma^2_2(x) \text{ for at least } 50\% \text{ of } x \in [0, 1],
\]

In this example, \( \varphi(x) = \sigma^2(x) \) and \( A = \{ \text{at least } 50\% \text{ of } x \in [0, 1] \} \).

Cox and Lee (2008) presented a method valid for testing the equality of two mean curves, assuming that they have identical covariance structures. Their hypothesis test is similar to the first example, i.e., in (4), except that the alternative hypothesis is stated as \( H_1 : \mu_1(x) \neq \mu_2(x) \) for at least one \( x \in [0, 1] \). Their method discretizes the continuous domain of the curve into a dense grid of points and performs univariate, pointwise \( t \) tests for two-sample mean comparisons. Cox and Lee showed that despite the pointwise tests, their method is in fact a functional comparison. Our test procedures borrow the ideas in Cox and Lee (2008) but need to expand the test statistics and adjust the permutation steps (discussed in Section 3.1.2).

3.1.1. Pointwise test

As Cox and Lee (2008) employ pointwise tests, they discretize the continuous domain of \( x \) to a grid of points, and let us denote the discretized domain by \( D_m \subset [0, 1] \). Accordingly, the discrete counterpart of \( A \) is denoted by \( A_m \subset D_m \). The pointwise tests of the original test in continuum are expressed as:

\[
H^o_0(x) : \varphi_1(x) = \varphi_2(x) \text{ for all } x \in D_m \\
\text{against } H^o_1(x) : \varphi_1(x) > \varphi_2(x) \text{ for all } x \in A_m.
\]

For the hypotheses in (6), individual univariate hypothesis tests are conducted to compare the two groups of curves evaluated at each input point \( x \). The individual test can be either a one-sided \( t \) test, for testing the inequality of the mean, i.e., when \( \varphi = \mu \), or a \( F \) test, for testing the inequality of the variance, i.e., when \( \varphi = \sigma^2 \). Each of the univariate two-sample tests computes a \( p \) value, denoted by \( p(x) \), at each grid point \( x \). Let us further denote the \( p \) values computed from the original data samples by \( p^o(x) \), where the superscript \( o \) indicates the original \( p \) values or observed \( p \) values. This superscript notation is introduced to differentiate the original \( p \) values from the \( p \) values generated from the re-sampled data. Recall that a smaller \( p \) value presents stronger evidence in favor of \( H_1(x) : \varphi_1(x) > \varphi_2(x) \).
3.1.2. Control familywise error rate

Using the pointwise tests, the comparison in continuum is approximated by a multiple comparison problem. We need to find a test statistic such that the family of the pointwise tests can be under the control of a prescribed significance level of $\alpha$. Analogous to the type-I error, the Family Wise Error Rate (FWER) is the probability of the false rejection of the complete null hypothesis that consists of multiple comparisons. That is, given $\alpha$,

$$Pr\{H_0^\text{w} \text{ is rejected} \mid H_0^\text{w} \text{ is true} \} \leq \alpha. \quad (7)$$

The conventional Bonferroni correction does not perform well for multiple comparisons involving functional data, because the events of rejecting the individual null hypotheses are highly correlated when the input $(x)$, with which the hypotheses are associated, are close. Then can one still control the FWER considering the inherent correlations in the functional data? The short answer is yes. The Westfall–Young permutation-based procedure (Westfall and Young, 1993) provides the ability to do so. Before diving into the specific procedure, we need to understand what we want to control, i.e., the test statistics.

**Test Statistics.** The choice of the test statistic for the purpose of FWER control depends on the subdomain of interest of $A_m$:

1. If $A_m = \{\text{at least one } x \in D_m\}$, the test statistic has been studied by Cox and Lee (2008), which is $\min p$, the minimal $p$-value among those computed from the family of individual tests. This is to say, as long as one $x$ exists, such that $\varphi_1(x) > \varphi_2(x)$ does not occur by chance, we accept $H_0^\text{w}(x)$. Equation (7) then becomes

$$Pr\{\min p \leq \alpha \} \leq \alpha. \quad (8)$$

2. If $A_m = \{\text{all } x \in D_m\}$, the test statistic is $\max p$, the maximal pointwise $p$-value. We control the probability of most probably event, that is, the probability of the individual test, $\varphi_1(x) > \varphi_2(x)$, evaluated at the $x$ that returns the maximal $p$-value among all $x \in A_m$. Should this individual test be under the significance level of $\alpha$, the $p$-values of every other event will be controlled as well. Equation (7) then becomes

$$Pr\{\max p \leq \alpha \} \leq \alpha. \quad (9)$$

3. If $A_m = \{\text{at least 50\% of } x \in D_m\}$, the test statistic is $\text{med} p$, the median of the pointwise $p$-values. We control the median pointwise $p$-value to be smaller than or equal to the significance level of $\alpha$, so that the events of $\varphi_1(x) > \varphi_2(x)$ associated with the $p$-values smaller than the median $p$-value will be controlled as well. Equation (7) then becomes

$$Pr\{\text{med} p \leq \alpha \} \leq \alpha. \quad (10)$$

We further denote the realization of $\min p$, $\text{med} p$ and $\max p$ by $\min p$, $\text{med} p$ and $\max p$, respectively.

**Procedure of the Test.** For generating the distribution of $\min p$, or $\max p$, or $\text{med} p$, Westfall and Young (1993) presented a permutation-based procedure. We use $\min p$ as an example to illustrate how to apply the Westfall–Young procedure to generate its distribution. For generating the distribution of $\max p$ or $\text{med} p$, simply change the boldface $\min$ in the following with either max or med, respectively.

Recall that given the two samples of curves, we calculate the pointwise $p$-values under the null hypothesis and record the minimal $p$-value as $p_{\min}^0$. We then randomly permute the curves between the two samples $N$ times. At each iteration, the two groups of resampling curves are generated. With the resampling data, we perform the pointwise tests and report the $p$-value of each individual test, denoted by $p_l(x)$, $l = 1, ..., N$, where the superscript $s$ indicates simulated $p$-values. Then, calculate $\min_{l \in A_s} p_l(x)$ and denote it by $p_{\min}^s$. Generate the empirical distribution of $\min p$ using $\{p_{\min}^l, l = 1, ..., N\}$. The corrected $p$-value with the FWER controlled for is $Pr\{\min p < p_{\min}^l\}$, that indicates the extent to which the observed familywise minimal $p$-value is extreme. This corrected $p$-value can be estimated by using the empirical distribution of $\min p$ obtained through the permutation procedure. Specifically, the estimate of the corrected $p$-value is through locating $p_{\min}^0 = \min_{l \in A_s} p_{\min}^s(x)$ on the empirical distribution of $\min p$. The detailed steps of the test procedure is presented in Algorithm 1.

**Algorithm 1** Estimate the $p$-value of the functional curve test

For $\text{stat} = \min$, $\max$ or $\text{med}$,

1. **Step 1.** Perform univariate one-sided $t$ tests for testing mean (or $F$ tests for testing variance) on $x_k \in A_m$, $k = 1, ..., m$. Compute the pointwise $p$-values, $p(x_k)$ and then $p_{\text{stat}} = \text{stat}\{p(x_k) : k = 1, ..., m\}$.

2. **Step 2.** Let $p_{\text{stat}}^0 \leftarrow p_{\text{stat}}$.

3. **Step 3.** For $l = 1, ..., N$, randomly permute the group label $t \in \{1, 2\}$ in the data $\{z_t(x_k) : j = 1, ..., J, k = 1, ..., m\}$. Repeat Step 1. Record $p_{\text{stat}}^l = p_{\text{stat}}$.

4. **Step 4.** Find $l_0$ such that $p_{\text{stat}}^{(1)} \leq \cdots \leq p_{\text{stat}}^{(l_0)} \leq \cdots \leq p_{\text{stat}}^{(N)}$, where $p_{\text{stat}}^{(k)} = p_{\text{stat}}$.

5. **Step 5.** The corrected $p$-value then is $\frac{l_0}{N}$.

We permute the whole curves between the two groups to preserve the inherent correlation among the points within a curve. This treatment is different from the practice in Cox and Lee (2008), in which the points are allowed to permute between the two groups. Their permutation does not cause any problem, as they assume a common covariance structure of the two groups, meaning that two samples of discrete points follow the same joint distribution. In our engineering decision process, such an assumption cannot be guaranteed. Permuting the curves relaxes such requirements, and doing so is also consistent with engineering practice, as an assembly of curve segments from different stages does not have a valid engineering meaning.

3.2. Hypothesis tests-based detection rule

As discussed in Section 1, when a polishing action improves the surface quality, one expects to notice either the mean
curve, especially its tail portions, is flattened, or the variance of the curves is reduced, or both. In other words, if any or all of the inequalities in (1) holds, it signals an improvement of the surface quality, suggesting that the polishing action is effective. Understandably, we would like to test the following three hypotheses on functional curves:

\[
\begin{align*}
H_0^{\mu_0}: & \mu_{l-1}(s) = \mu_l(s) \quad \text{for all } s \in [0, \tau] \\
H_1^{\mu_0}: & \mu_{l-1}(s) > \mu_l(s) \quad \text{for all } s \in [0, \tau] \\
H_0^{\alpha_0}: & \mu_{l-1}(s) = \mu_l(s) \quad \text{for all } s \in [1-\tau, 1] \\
H_1^{\alpha_0}: & \mu_{l-1}(s) < \mu_l(s) \quad \text{for all } s \in [1-\tau, 1] \\
H_0^{\sigma^2_0}: & \sigma_{l-1}^2(s) = \sigma_l^2(s) \quad \text{for all } s \in [0, 1] \\
H_1^{\sigma^2_0}: & \sigma_{l-1}^2(s) > \sigma_l^2(s) \quad \text{for at least 50% of } s \in [0, 1].
\end{align*}
\]

The superscripts, $\mu_0$ and $\alpha_0$, indicate that the respective hypothesis is tested for the upper tail or lower tail of the mean functions, and the superscript, $\sigma^2$, indicates the variance function test. These hypothesis tests have clear physical interpretation. The alternative hypothesis in (11) suggests that the peaks are being flattened; the alternative hypothesis in (12) suggests that the valleys are being filled; and the alternative hypothesis in (13) means that the surface is getting more even. A null hypothesis is rejected when there is strong evidence against it. The strength of the evidence is quantified by the $p$-value; the smaller, the stronger.

To perform the pointwise tests, we generate a grid of evaluation points $D_m = \{s_1, \ldots, s_m\} \subset [0, 1]$. Evaluate every function $\varphi_j(s)$ at these finite points, yielding the data vectors $\varphi = (\varphi_j(s_1), \ldots, \varphi_j(s_m))^T$. Define $D_u = \{s_k, k = 1, \ldots, u : s_k \leq \tau, s_k \in D_m\}$ and $D_l = \{s_k, k = m - l, \ldots, m : s_k \geq 1-\tau, s_k \in D_m\}$. Upon discretizing the continuum of comparisons, the three functional hypotheses are approximated with three families of individual univariate hypotheses:

\[
\begin{align*}
H_0^{\mu_0} : & \mu_{l-1}(s) = \mu_l(s) \quad \text{for all } s \in D_u \subset [0, \tau], \\
H_1^{\mu_0} : & \mu_{l-1}(s) > \mu_l(s) \quad \text{for all } s \in D_u \subset [0, \tau],
\end{align*}
\]

\[
\begin{align*}
H_0^{\sigma^2_0} : & \sigma_{l-1}^2(s) = \sigma_l^2(s) \quad \text{for all } s \in D_u \subset [0, 1], \\
H_1^{\sigma^2_0} : & \sigma_{l-1}^2(s) > \sigma_l^2(s) \quad \text{for at least 50% of } s \in D_u \subset [0, 1].
\end{align*}
\]

We control the FWER of the hypothesis in (14) and (15) using the test statistic $\max P$ as in (9), and control the FWER of the hypothesis in (16) using the test statistic $\text{med} P$ as in (10). The corresponding testing procedure is described in Section 3.1.2.

Combining the three tests, our rule for surface change detection is—if the null hypothesis of any of the three sub-families is rejected, it signals that the current polishing action is effective and should be continued; otherwise, the polishing action did not make statistically significant change to the surface under polishing. To control the type-I error of this final detection rule, we invoke Bonferroni’s method for multiple comparisons. The three families of tests, as in (14)-(16), can be considered independent of each other, so that Bonferroni’s method sets the type-I error rate for each of the test families to be less than or equal to $\alpha/3$. Within each family, these pointwise tests are highly correlated, so that the respective FWER is controlled through the Westfall–Young procedure, as outlined in Algorithm 1.

Please note that the above decision rule is for detection of a surface change. Once a surface change is not detected, how engineers should react to it needs further consideration. Generally speaking, there are two options: (i) cleaning
the polishing tool or (ii) changing to a finer tool. The second option also leads to a stopping decision of the overall post-processing, if the finest polishing tool has already been used.

Between the choice of cleaning the tool and changing to a finer tool, the decision is usually not hard. It does not cost much to clean the tool and use the clean tool to polish a little further. When there is still no improvement in the surface quality, it is then clear that a new tool is needed, or if no finer tool is there to be used, the polishing process should naturally be stopped at that point.

4. Physical experiments

We apply the proposed hypothesis test-based detection method to the polishing/lapping process of the spherical surface of silicon bead (ball), introduced in Section 1. To achieve its nano-scale finish, the coated spherical bead is passed through the Tip Truncating (TT) process, the lapping process and the polishing process; the tools are refined as the beads go through these different processes. Each type of processes is divided into several stages, and thus the surface roughness at the corresponding intermediate stages can be measured. See Figure 2 for the illustration of the arrangement of the surface roughness data at a process stage.

Figure 4(a) demonstrates the data configuration of the surface roughness in a matrix of Z, where the pixel height is represented by \([z_{w,v}]\) and \(w, v\) are the pixel position indices along the two directions. The \(X\) and \(Y\) coordinates are denoted by \(X_w\) and \(Y_v\), respectively. The pixels are spread evenly over the location, with a between-pixel distance of 0.359 \(\mu m\) along the \(X\) coordinate and 0.369 \(\mu m\) along the \(Y\) coordinate.

4.1. Pixel height calibration

The original pixel height matrix of \(Z\) embodies the systematic local surface trend more than the pixel heights relative to its local surface baseline. With the local surface trend included, one can see in Figure 5(a) that the rough portion of the pixel heights can be overwhelming (everything appears as bright white).

Considering the small size of the local area (roughly 0.2 \(\times\) 0.17 mm) associated with each sample location, it is reasonable to approximate the systematic local surface trend with a sphere, regardless of the overall geometry of the product. Our data pre-processing consists in calibrating the pixel heights by removing the systematic local trend. This estimation of the local trend is also known as "soft-fixturing" in metrology and is commonly used in other similar studies; please see, for example, Figure 3 in Xia et al. (2011), that illustrates different surfaces and measurements. In their figure, the systematic local surface corresponds to the “dashed-line geometry”, and taking the difference between the dashed-line geometry and the measured pixel heights calibrates the pixel heights to be relative to the local surface baseline.

As our product happens to be a sphere, it may leave readers with a confused idea that what we do here is to flatten the overall geometry of the product. That is not true. What we do is limited to a small area on the surface (basically, a dot) for the purpose of calibrating the pixel height. The global curvature, or whatever the complexity of the overall product’s geometry, is not affected by this data pre-processing action. The following steps outline the process for pixel height calibration:

1. Convert the matrix \(Z\) to a vector \(z\). We first read the matrix \(Z\) row-wise into a row vector of size \(I \times 1\); see Figure 4(b). For the problem at hand, \(I = 640 \times 480 = 307,200\). Denote the \(i\)th element of \(z\) by \(z_{(i)}, i = 1, ..., I\). We also re-index \(X_w\) and \(Y_v\) to become \(X_{(i)}\) and \(Y_{(i)}\) respectively. We want to ensure that \(z_{(i)}\) is associated with its original coordinates, i.e., those with which \(z_{w,v}\) is associated. To accomplish this, we let \(X_{(i)} = X_w\) and \(Y_{(i)} = Y_v\), where \(w = \left\lfloor \frac{i - 1}{480} \right\rfloor\) and \(v = i \mod 480\). For example, for the pixel height of \(z_{3240}\) in the original \(Z\) matrix, it is the 1200th element in the vector of \(z\). Given \(i = 1200\), the index relationship explained above is then \(w = \lfloor 1200/480 \rfloor = 3\) and \(v = 1200 \mod 480 = 240\), which connects \(z_{(1200)}\) with \(z_{3240}\).

2. Arrange the three newly created row vectors, \(X\), \(Y\) and \(z\) into a matrix as \([X^T, Y^T, z^T]\), where \(X = [1, ..., X_{(i)}]\), \(Y = [1, ..., Y_{(i)}]\), and \(z = [1, ..., z_{(i)}]\). Compute the coordinates of the center of the sphere, denoted by \([X_c, Y_c, z_c]\) and the radius of the sphere, denoted by \(r\). Such computation can be facilitated by using some software routines, such as the MATLAB® built-in function sphereFit.
3. Subtract the surface baseline from the vector \( z \). Calculate the new pixel height, \( z_0 \), such that

\[
z_0(X_w, Y_v) = \sqrt{(r^2 - (X_w - X_c)^2 - (Y_v - Y_c)^2)} + z_c.
\]

The new pixel height is calibrated by subtracting the adjusted surface baseline from the original measurements.

4. The new pixel height is calibrated by subtracting the adjusted surface baseline from the original measurements. Figure 5(b) displays the surface roughness after the pixel height calibration. The rough texture of the surface is much more visible than the plot on the left. For the sake of notational simplicity and without ambiguity, we still use \( z \) instead of \( z_0 \) to represent the pixel height after the calibration.

4.2. Study of the bead polishing process

We have conducted two experiments to illustrate the use of the method for detecting the lack-of-improvement point during a polishing process and then a lapping process. This subsection focuses on the polishing process in the first experiment, whereas Section 4.3 presents the second experiment, which has only a lapping action. Part of the reason that we skip the analysis of lapping in the first experiment is due to insufficient locations sampled on the bead at the lapping stages for our permutation-based method to be applied.

The first experiment comprises of a stage of TT, four stages of lapping, and nine stages of polishing. Figure 6(a) presents the trend of \( S_d \) from coating to the last polishing stage. The notation along the horizontal axis indicates that the \( S_d \) displayed is measured right after that specific stage. For instance, “Lap 3” means the third time when the lapping process was paused and the measurement of roughness was taken. Figure 6(b) presents a zoom-in version of the \( S_d \) box-plots for the nine polishing stages.

The \( S_d \) value of the surface is 453.2 nm after coating and reduced to 395.7 nm by the TT process. Afterward, the four stages of lapping process bring \( S_d \) significantly down to an average of 24.8 nm, suggesting that the lapping process makes significant progress in smoothing the initial rough surface. Up to this point, the use of \( S_d \) is adequate and there is hardly any disagreement on decision making or action taking.

From the first stage of the polishing process and onward, however, the reduction in \( S_d \) is not that significant. The zoom-in view on Figure 6(b) shows some degree of further reduction, from roughly 12.5 nm to a little bit over 10 nm. However, the cost to accomplish that reduction is huge. From Polish 1 to Polish 9, the time spend is a total of 240 hours, or 10 days. Under such a long time of polishing, the risk has increased considerably that the bead, which is hollow inside, could crack, or other damages may happen to it. It would be preferred if such overpolishing can be avoided by triggering an earlier stopping point.

4.2.1. Hypothesis tests results versus \( S_d \)

Table 1 presents the results obtained using the hypothesis testing-based decision process. It includes the \( p \)-values associated with the three hypothesis tests and the decisions that the \( p \)-value suggests. Recall that the detection rule is that as long as any of the null hypotheses is rejected, the surface quality is considered to be improved beyond random fluctuation. If none of the null hypotheses is rejected, we deem that no significant surface quality improvement has been detected.

To register a decision, a threshold for the \( p \)-value is needed. However, the \( p \)-value thresholds chosen in practice, for instance, the customary 0.05 cut-off, bears a certain degree of arbitrariness. A closer look at the \( p \)-values in Table 1 offers a clue in selection. It is apparent that the \( p \)-values therein belong to three groups, \{0.02, 0.04, 0.05, 0.06\}, \{0.10, 0.12\}, and \{> 0.22\}. Recall that we have three independent hypothesis test families in each decision. By invoking Bonferroni adjustment, the type-I error for the overall decision is three times of that of each family. Should we deem a \( p \)-value of 0.10 or 0.12 significant for each family, it implies the type-I error for the overall decision is at the level of 0.30 or greater.
The hypothesis testing outcomes shown in Table 1 detect the variance reduction from Stage P3 to Stage P4, and to a lesser degree of significance, a lowered upper tail from Stage P5 to Stage P6, a reduced variance from Stage P6 to Stage P7 and a lowered upper tail from Stage P8 to Stage P9. These detection outcomes show a good consistency with observing the Sa boxplots in Figure 6(b). The detection of reduced variance is more consistent between the hypothesis tests and the Sa boxplots. The hypothesis tests signal two instances of reduced variance, from Stage P3 to Stage P4 and then from Stage P6 to Stage P7, both of which are visually so in the Sa boxplots. Even though these outcomes are consistent, the merit of using the hypothesis testing method is that the new method provides a more detailed information that can be used to inform users about where a change is detected and to what degree it is significant.

On the other hand, not all the detection outcomes using the hypothesis testing method are the same as using Sa. Using the Sa values appears to be more ready in signaling an improvement in the surface quality. Consider the decreasing trend of Sa from Stage P3 to P6 in Figure 6(b). By contrast, the hypothesis tests do not detect significant improvement on the surface after these polishing stages, per the criteria defined in (14)–(16)), except that some marginal improvement at the upper tail (i.e., peak removals) from P5 to P6.

In Figure 7, we plot the average BACs and the associated 96.7% confidence bands (96.7% corresponds to the 0.33 p-value cut-off) on two consecutive stages to visualize the two-stage surface roughness changes and discern if the outcome difference between the hypothesis tests and Sa makes sense. Since the median Sa shows a decreasing trend, we hope to see that: (i) the mean curve of Stage t (solid red curve) representing the peaks (Quantile 0 to 0.25) is lowered compared with that of Stage t – 1 (solid blue curve); and/or (ii) the mean curve of Stage t, representing the valleys (Quantile 0.75 to 1) is raised. The reality is that some of the tail portions of the BACs do not show significant differences. Even if there are some differences, those appear to be well within the confidence bands. We mark certain areas in Figure 7 with dashed circles to highlight. It is evident to us that the median Sa fails to capture such regional differences, but these regional differences are crucial in reaching a sensible decision for these ultra-precision manufacturing products.

We also observe that once in a while, the BACs at a later stage of polishing could deteriorate as compared with the preceding stage. Consider the example of the lower tail portion from Stage P5 to Stage P6 (Figure 7 (c)), representing their valleys. It turns out that after additional polishing the valleys at Stage P6 are deeper than those at Stage P5, mostly likely due to extra scratches introduced on the surface during polishing, which is a clear indication of over polishing. Please note that in our plot, we only plot the Quantile 0.75 to 0.998 while leaving out a tiny portion close to one, because that extremely deep valley pixels are sometimes of much great depth and their presence compresses the scale on the plot, making the visualization difficult.

If using the hypothesis testing outcomes in Table 1, the polishing process should stop at Stage P2, or if not so quickly, at Stage P5 as the latest. Doing either will save significant polishing time, reduce materials removal, and lower the risk of over polishing and bead damaging in the long operation stretch.

### 4.3. Study of the bead lapping process

The second experiment comprises a stage of coating, a stage of TT and 10 stages of lapping. Figure 8(a) presents the boxplots of Sa for the whole process, whereas Figure 8(b) is the zoom-in view of Stage L6 and onwards. The lapping process, after L5, is able to bring the surface roughness to the level of 12 nm. This time, the pattern of fluctuation in Sa is much more obvious, making the decision harder using the traditional decision tools. From Stage L6 and onward, the time spent, a total of 80 hours, and the material removed, 7.72 mg, are again significant.

The hypothesis tests results shown in Table 2 detect the variance reduction from Stage L7 to L8, but do not confirm the mean reduction trend as observed in the Sa plot for Stage L8 to L10, except for a marginal peak flattening from L8 to L9. Inspecting the Sa boxplots, the variance reduction from Stage L7 to L8 is not so obvious. We therefore present the detailed BACs and the associated confidence bands in

| Stages | Mean test for upper tail | Mean test for lower tail | Variance test | Surface quality improvement detection |
|--------|--------------------------|--------------------------|---------------|----------------------------------------|
|        | p-value                  | Outcome                  | p-value       | Outcome                               |
| P1 vs. P2 | 0.58                    | Not lowered              | 0.22          | Not raised                            |
| P2 vs. P3 | 0.89                    | Not lowered              | 0.94          | Not raised                            |
| P3 vs. P4 | 0.65                    | Not lowered              | 0.98          | Not raised                            |
| P4 vs. P5 | 0.10                    | Not lowered              | 0.27          | Not raised                            |
| P5 vs. P6 | 0.06                    | Lowered (marginal)       | 0.28          | Not raised                            |
| P6 vs. P7 | 0.12                    | Not lowered              | 0.74          | Not raised                            |
| P7 vs. P8 | 0.96                    | Not lowered              | 0.91          | Not raised                            |
| P8 vs. P9 | 0.04                    | Lowered (marginal)       | 0.81          | Not raised                            |

# in the Stages column indicates a specific polishing stage.
Figure 7. Mean curves (solid curves) and 96.7% confidence bands (dashed curves) of BACs at two consecutive stages: (a) Polish 3 vs. Polish 4; (b) Polish 4 vs. Polish 5 and (c) Polish 5 vs. Polish 6.
Figure 9 for Stage L7 (blue curves) versus L8 (red curves). On these plots, it is clearly showing that the band of L8 is narrower than that of L7.

Figure 10 presents the BACs of Stage L6 and onward and of each two consecutive stages being compared. The mean curve comparison show that between Stages L6 and L7, the reason for a lack of detection by the hypothesis testing method is due to the highest peak and the deepest valley not being improved. Some may argue that those are outliers and should not be considered, whereas others may argue that the extreme of the peaks and valleys does reflect the surface roughness, as they reveal the bumps and scratches produced during the action of lapping. We want to note that our decision process can be easily tailored to suit different needs in the specific contexts. In the case where the sensitivity to the peaks and valleys is considered the above results show that the hypothesis tests do have the desired sensitivity in detection. If the highest peaks or the deepest valleys should be excluded from decision making, one just needs to adjust the definition of $D_u$ and $D_l$ in (14)–(16) to accommodate such changes. The rest of the testing procedure stays more or less the same.

### Table 2. Test results of L6 and onward. L# in the Stages column indicates a specific lapping stage.

| Stages       | Mean test for upper tail | Mean test for lower tail | Variance test | Surface quality improvement detection |
|--------------|--------------------------|--------------------------|---------------|---------------------------------------|
| L6 vs. L7    | p-value | Not lowered | p-value | Not raised | p-value | Not reduced | No improvement detected |
| L7 vs. L8    | 0.99     | Not lowered | 1       | Not raised | 2 x 10^-5 | Reduced | Improvement detected |
| L8 vs. L9    | 0.06     | Lowered (marginal) | 0.94 | Not raised | 0.83 | Not reduced | Improvement detected (marginal) |
| L9 vs. L10   | 0.41     | Not lowered | 0.99 | Not raised | 0.83 | Not reduced | No improvement detected |

Figure 8. $S_a$ boxplots for the second physical experiment. (a) For all stages; (b) for Stage L6 and onward.
4.4. Detection sensitivity analysis

In this section, we perform sensitivity studies on the number of sampling locations in Section 4.4.1 and on the value of \( \tau \) in Section 4.4.2. Recall that \( \tau \) specifies the tail portion used in (14)–(16).

4.4.1. Sensitivity on number of sampling locations

The experiments leading to the hypothesis test results in Tables 1 and 2 are based on nine sampling locations, randomly sampled on the bead’s surface. When there are too few curves, i.e., too few sampling locations, a permutation tends to repeat the combinations and thus becomes less effective. On the other hand, doing on-product measurements is expensive, especially on a small product such as the beads in our process. Thus, operators do not find it affordable to sample too many locations. The purpose of our sensitivity study here is to see whether a sample size of nine locations can lead to consistent conclusions.

As lapping and polishing are disruptive operations, once utilized, the surface cannot be restored to its original state to take more measurements. To support a sensitivity study, we therefore conduct a new experiment with more sampling locations taken during the process. This is a lapping process with five operating stages, labeled as L1 through L5. Please note that because this is a brand-new experiment, these

Figure 10. Mean curves (solid curves) and 6\( r \) confidence bands (dashed curves) of the BACs for two consecutive stages for Stage L6 and onward: (a) Lap 6 vs. Lap 7; (b) Lap 8 vs. Lap 9; (c) Lap 9 vs. Lap 10.
lapping stages are not directly comparable to the lapping or polishing stages in the earlier sections. For each operating stage, we divide the bead surface into three shells and randomly sample at three, four, and five locations on each shell, respectively. In total, the number of sampling locations are 9, 12 and 15 locations, respectively. The results using different sampling locations are shown in Table 3.

We make two observations. The first observation is that the overall conclusion is reasonably consistent and stable when using nine or more sampling locations—other than the variance test of Stage L2 vs. L3, all other tests lead to the same decision. The second observation is that using more sampling locations does help improve the robustness of the decision, but, of course, at a higher measurement cost. We note that the decision process based on 12 and 15 sampling locations reached an opposite conclusion in the variance test of L2 vs. L3. Apparently when the number of sampling locations increases, the initial difference between the variance curves is reduced, to the degree that it cannot reject the null hypothesis. Please see the variance curves on the two stages shown in Figure 11.

Through this analysis, our recommendation is for the operator to use a slightly larger number of sampling locations if affordable. But in consideration of economic operation, nine sampling locations are deemed an acceptable practice, especially considering that using nine sampling locations yields a conservative decision, which is to continue polishing, rather than stopping prematurely. We also confirm that when analyzing this new set of data with more sampling locations, we did not come across any violations or contradictions of the messages obtained through the previous nine-location experiments.

4.4.2. Sensitivity on $s$

Our default choice is $s = 25\%$. Intuitively, this $s$ choice says that the top quarter is the upper tail, the bottom quarter is the lower tail, and the 50% in between is the middle body. We consider that this is a reasonable choice, allocating a sufficiently large portion to cover the two tails. For both 9 and 15 sampling locations, we here conduct another sensitivity analysis by varying the tail portion from 5% to 33%, i.e., by comparing outcomes of six tail-portion settings, i.e., $s = 5\%$, 10%, 15%, 20%, 25% and 33%, respectively, and see whether the decision outcomes are sensitive to the choice of the tail portion. The outcomes of the mean tests are listed in Table 4. We did not notice any significant difference in the p-values, and certainly no change in the detection outcomes associated with any one of the hypotheses.
4.5. Test on a flat surface

As explained earlier, a main advantage of the hypothesis test-based method is that it is applicable to both flat and non-flat surfaces. In this subsection, we show its applicability to the flat surface studied in Jin et al. (2020) using the data from their paper. For their polishing experiment details, please refer to Jin et al. (2020).

Table 5 summarizes the test results. The right half of Table 5 presents the p-values of the three hypothesis tests: the upper tail mean test, the lower tail mean test, and the variance test. Consistent with studies presented in the preceding subsections, the p-value cut-off used is 0.1 at the overall decision level, or 0.033 at the individual test level. Should any of the three null hypotheses be rejected, a change in surface quality is considered to have been detected and the action suggested is to keep polishing with the current tool. Otherwise, no surface change is detected and the action suggestion is to change the current polishing tool or stop altogether.

The column “GPBD-guided action suggestion” shows the suggestions given by the GPBD proposed by Jin et al. (2020). The GP-guided decision process is to compare the similarity between sampling locations, quantified by the scale parameter in the GP model, between Stage $t-1$ and Stage $t$. Jin et al. (2020) deem the across-stage similarity a
good proxy to inform about the roughness of the surface. The decision is then made by comparing the change in the scale parameter in the GP models associated with the two consecutive stages.

By comparing the decision outcomes from the two methods, we notice that every time when GPBD suggests a change, the HT-based method also suggests the same. On the other hand, the HT-based method suggests more changes, and more importantly, earlier change points. Should the HT-based method be followed, the first change action would have been at Stage 6, five stages earlier than that suggested by GPBD. This suggestion means that one ought to change the polishing tool from its initial 800-grit pad to a finer 1200-grit pad. By looking at the value of \( \hat{\sigma} \), we can see the merit of this new suggestion. From Stage 6 through Stage 11, \( \hat{\sigma} \) fluctuates between 0.135 \( \mu \)m and 0.174 \( \mu \)m with no clear sign of reducing. It appears that using the 800-grit pad was effective to bring down the surface roughness from a very rough raw surface to the level of 0.150 \( \mu \)m, namely 150 nm, but was ineffective in making further inroads. This may be the result of the inherent capability of the 800-grit pad. Had one changed the pad after Stage 6, the next two changes suggested by the HT-based method may not necessarily happen anymore, because once a new pad is used, the product surface would react differently and the change in the surface roughness usually follows a different course.

From the method design point of view, we can also see that the HT-based decision process presents certain advantages over the GPBD-guided decisions. As explained above, GPBD characterizes the spatial correlation between sampling locations; the stronger the correlation, the smoother the surface. This is reasonable, but the information in the roughness is aggregated into a scalar correlation parameter. Jin et al. (2020) demonstrated that this scalar correlation parameter is still more informative than the median \( Sa \) value. Yet, in the process of condensing the information associated with BACs into scale parameters, certain fine granularity of information could get lost. By contrast, the HT-based method detects a surface quality change by comparing two sets of curves without the data compression steps, so it appears more robust.

5. Concluding remarks

We propose in this article a method for enabling endpoint decisions in surface finishing processes of precision manufacturing products. Our method does not require the product geometry to be a flat surface. In fact, we illustrate its applicability to the polishing/lapping of spherical beads. The method detects the surface quality change by comparing the clusters of BACs from two consecutive polishing stages. When no change is detected, it signals the need for the polishing action to change or stop.

There are a number of unique features of the method worth highlighting. Compared with GP-based or other model-based methods, the proposed decision method does not need to specify an underlying model and rely on strong assumptions. The proposed method is built upon statistical testing of nonparametric quantile curves, which are a direct result of sorting the measurements of surface roughness. Such quantile curves, known as the BACs in engineering, are readily available in many different kinds of finishing processes. This explains why the method can be applied to both lapping and polishing processes, both spherical and flat surfaces, without changes in its underlying procedure. The method is able to test the relationship of quantile curves for a specified region, or even, a union of disconnected sub-regions, should prior physical knowledge advise such choices. Compared with a surface roughness measure such as \( Sa \), the proposed decision method not only reveals the detailed nature of a change in surface roughness, but also informs the statistical significance of such change.

Adopting the proposed decision process leads to earlier stopping or tool-changing actions, which could save a lot of time, energy, or material removal, without sacrificing the quality of the final surface finish. More than that, shortening the final long stretch of polishing has the added benefit of reducing the risk that the polishing products may become cracked, scratched, or otherwise damaged.

Funding

The authors acknowledge the generous support from National Science Foundation under grant no. IIS-1849085, Department of Energy grant no. DE-AC52-07NA27344 through Subcontract no. B646055 and B640459 via the Lawrence Livermore National Laboratory, and Texas A&M Office of President’s X-grant Program.

Notes on contributors

Shilan Jin is currently a PhD candidate in the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX. She holds a BSc in mathematics from Jilin University, China and a M.Sc. in industrial and systems engineering from Texas A&M University. Before pursuing her PhD, she was a Decision Scientist with the Walt Disney Company, Orlando, FL, from 2013 to 2017. Her research interests include data analytics, machine learning, and statistical modeling applied to advanced manufacturing and materials research.

Rui Tuo received his PhD in statistics at University of Chinese Academy of Sciences in 2013. Currently, he is an assistant professor in the Department of Industrial and Systems Engineering and an adjunct assistant professor in the Department of Statistics at Texas A&M University. Tuo’s research interest mainly lies in data science for computer simulations, including design and analysis of computer experiments and uncertainty quantification. Recently, he has also been working on theory and computational techniques for nonparametric models and Gaussian process regression.

Akash Tiwari received his BTech in industrial and systems engineering from the Indian Institute of Technology, Kharagpur, India, in 2019. He is currently working towards a PhD degree with the Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, USA. His research interests include cybersecurity and data analytics for manufacturing. He was a Summer Intern with the Royal Enfield Motors Factory, Chennai, India, in 2017. In 2018, he was a Summer Research Intern with the Durham University Business School, Durham, U.K. He is a student member of the Institute for Operations Research and the Management Sciences.

Satish Bukkapatna received his PhD degree in industrial and manufacturing engineering from Pennsylvania State University (1997). He is currently the Rockwell International Professor of Industrial & Systems Engineering at Texas A&M University, and the Director of Texas A&M Engineering Experiment Station Institute of Manufacturing Systems. His research interests are broadly in smart manufacturing.
systems, and ultraprecision manufacturing. Dr. Bukkapatnam is a Fellow of IIE and SME.

Chantel Aracne-Ruddle conducts research at Lawrence Livermore National Laboratory. During her tenure she has worked on multiple, wide-ranging projects, including the development of techniques for polishing single crystal materials to be studied at static pressures in the megabar range. She is currently the manager for the Target Materials Science and Technology Laboratory for the National Ignition Facility (NIF). Her current focus is developing sphere polishing techniques that minimize roughness while controlling sphericity and concentricity. Her focus interests are down-selection of materials, their qualification, and the development of techniques to assemble them as micrometric components to support experiments at the NIF.

Ariel M. Lighty received her BSc in chemical engineering from the University of Rochester. She was previously a research assistant at the Laboratory for Laser Energetics and then Lawrence Livermore National Laboratory. She is currently a chemical and biomolecular engineering PhD student and Arthur A. Schomburg Fellow at the University at Buffalo.

Haley Hamza is a current master’s student in the Graduate School of Arts and Sciences at Georgetown University. She holds a BS in biochemistry from University of California at San Diego. Her research interests are in atmospheric chemistry, particularly in measuring the downstream effects of carbon-based particulate matter from wildfire events.

Yu Ding received his PhD degree in mechanical engineering from University of Michigan (2001). He is currently the Mike and Sugar Barnes Professor of Industrial & Systems Engineering, Professor of Electrical & Computer Engineering at Texas A&M University, and Associate Director for Research Engagement at Texas A&M Institute of Data Science. His research interests are in data and quality science. Dr. Ding is a Fellow of IIE and ASME and serves as the Editor in Chief of IIE Transactions for the term of 2021-2024.

References

Benko, M., Härdle, W. and Kneip, A. (2009) Common functional principal components. The Annals of Statistics, 37(1), 1–34.

Biener, J., Ho, D., Wild, C., Woerner, E., Biener, M., El-dasher, B., Hicks, D., Egbert, J., Celliers, P., Collins, G., Teslich N., Jr, Kozioziemski, B., Haan, S. and Hamza, A. (2009) Diamond spheres for inertial confinement fusion. Nuclear Fusion, 49(11), 112001 1–5.

Bukkapatnam, S.T., Iquebal, A.S. and Kumara, S.R. (2018) Planar random graph representations of spatiotemporal surface morphology: Application to finishing of 3-d printed components. CIRP Annals, 67(1), 495–498.

Cox, D.D. and Lee, J.S. (2008) Pointwise testing with functional data using the Westfall–Young randomization method. Biometrika, 95(3), 621–634.

DeJure, R. (1997) CMP challenges below a quarter micron. Semiconductor International, 20(13), 54–58.

Frazier, W.E. (2014) Metal additive manufacturing: A review. Journal of Materials Engineering and Performance, 23(6), 1917–1928.

Fremdt, S., Steinebach, J.G., Horváth, L. and Kokoszka, P. (2013) Testing the equality of covariance operators in functional samples. Scandinavian Journal of Statistics, 40(1), 138–152.

Ganti, S. and Bhushan, B. (1995) Generalized fractal analysis and its applications to engineering surfaces. Wear, 180(1-2), 17–34.

Hall, P. and Hart, J.D. (1990) Bootstrap test for difference between means in nonparametric regression. Journal of the American Statistical Association, 85(412), 1039–1049.

Iquebal, A.S., El Amri, S., Shrestha, S., Wang, Z., Manoghran, G.P. and Bukkapatnam, S. (2017) Longitudinal milling and fine abrasive finishing operations to improve surface integrity of metal am components. Procedia Manufacturing, 10, 990–996.

ISO 4287 (1997) Geometrical Product Specifications (GPS), Surface Texture: Profile Method–Terms, Definitions and Surface Texture Parameters. International Organization for Standardization, Geneva, Switzerland.

Jiang, X. (2012) Precision surface measurement. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 370(1973), 4089–4114.

Jiang, X., Scott, P.J., Whitehouse, D. and Blunt, L. (2007a) Paradigm shifts in surface metrology. Part I. Historical philosophy. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 463(2085), 2049–2070.

Jiang, X., Scott, P.J., Whitehouse, D.J. and Blunt, L. (2007b) Paradigm shifts in surface metrology. Part II. The current shift. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 463(2085), 2071–2099.

Jin, S., Iquebal, A., Bukkapatnam, S., Gaynor, A. and Ding, Y. (2020) A Gaussian process model-guided surface polishing process in additive manufacturing. Journal of Manufacturing Science and Engineering, 142(1), 011003 1–12.

King, E., Hart, J.D. and Wehrly, T.E. (1991) Testing the equality of two regression curves using linear smoothers. Statistics & Probability Letters, 12(3), 239–247.

Kulasekera, K. (1995) Comparison of regression curves using quasi-residuals. Journal of the American Statistical Association, 90(431), 1085–1093.

Liu, J.P., Beyca, O.F., Rao, P.K., Kong, Z.J. and Bukkapatnam, S.T. (2017) Dirichlet process Gaussian mixture models for real-time monitoring and their application to chemical mechanical planarization. IEEE Transactions on Automation Science and Engineering, 14(1), 208–221.

Munk, A. and Dette, H. (1998) Nonparametric comparison of several regression functions: Exact and asymptotic theory. The Annals of Statistics, 26(6), 2339–2368.

Neumeyer, N. and Dette, H. (2003) Nonparametric comparison of regression curves: An empirical process approach. The Annals of Statistics, 31(3), 880–920.

Panaretos, V.M., Kraus, D. and Maddocks, J.H. (2010) Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. Journal of the American Statistical Association, 105(490), 670–682.

Prakash, A., Tuo, R. and Ding, Y. (2022). Gaussian process-aided function comparison using noisy scattered data. Technometrics, 64(1), 92–102.

Rao, P.K., Beyca, O.F., Kong, Z., Bukkapatnam, S.T., Case, K.E. and Komanduri, R. (2015) A graph-theoretic approach for quantification of surface morphology variation and its application to chemical mechanical planarization process. IIE Transactions, 47(10), 1088–1111.

Stewart, M. (2000) A new approach to the use of bearing area curve. Society of Manufacturing Engineers. Available http://numericalengineering.com/wp-content/uploads/2018/03/Training-A-New-Approach-to-the-Use-of-Bearing-Area-Curve-FC900229.pdf. (accessed 24 December 2020).

Wang, Z., Bukkapatnam, S.T., Kumara, S.R., Kong, Z. and Katz, Z. (2014) Change detection in precision manufacturing processes under transient conditions. IIE Transactions, 46(1), 449–452.

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

Whitehouse, D. (1982) The parameter rash—is there a cure? Wear, 83(1), 75–78.

Whitehouse, D. (2001) Fractal or fiction. Wear, 249(5-6), 345–353.

Xia, H., Ding, Y. and Mallick, B.K. (2011) Bayesian hierarchical model for combining misaligned two-resolution metrology data. IIE Transactions, 43(4), 242–258.

Zhang, J.T. (2013) Analysis of Variance for Functional Data. CRC, Boca Raton, FL.