Disease Markers 20 (2004) 295–307
IOS Press

Measurement reproducibility in the early stages of biomarker development

Walter S. Liggett[a,*], Peter E. Barker[a], O. John Semmes[b] and Lisa H. Cazares[b]
[a]*Statistical Engineering Division and Biotechnology Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA*
[b]*Center for Biomedical Proteomics, Eastern Virginia Medical School, Norfolk, VA 23507, USA*

Abstract. Biomarker discovery and development requires measurement reproducibility studies in addition to case-control studies. Parallel pursuit of reproducibility studies is especially important for emerging technologies such as protein biomarkers based on time-of-flight mass spectrometry, the case considered in this paper. For parallel studies, a way to improve reproducibility prior to identification of protein species is necessary. One approach is use of functional principal components analysis (PCA) as the basis for assessing measurement reproducibility. Reproducibility studies involve repeated measurement of a reference material such as a human serum standard. Measurement in our example is by SELDI-TOF (surface-enhanced laser desorption and ionization time-of-flight) mass spectrometry. Reproducibility is defined in reference to a source of variation, which in our example is associated with a type of commercially available protein biochip. We obtained spectra for 8 spots on each 11 chips. Two spectra are generally more alike when obtained from the same chip rather than different chips. Thus, our experiment indicates potential improvements from reducing variation in chip manufacture and chip handling during measurement. Our analysis involves careful registration of the spectra and characterization of the spectral differences. As shown by our example, a metrological analysis may enhance case-control studies by guiding optimization of the measurements underlying the biomarker.

1. Introduction

A biomarker is a biologically-based surrogate feature whose quantification reproducibly and reliably predicts or defines a broader, more comprehensive biological response that is too complex, impractical or technically impossible to quantify directly. As the Biomarkers Definitions Working Group [6] writes,

> "Biological marker (biomarker): A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention."

Development of a biomarker involves both measurement and interpretation [4]. Each of these aspects should be fully explored when advances in analytical instrumentation offer opportunities for the measurement of new biomarkers.

The phrase "objectively measured" implies that there is a measurement protocol that is followed each time a measurement is made. The goal of measurement system development is optimization of the protocol. An important aspect of the protocol is the variation encountered as the protocol is followed from one time to the next. Experimental investigation of this aspect is based on measurement of the same specimen more than once under different realizations of the protocol, that is, different ways of following the protocol. What constitutes different realizations changes with the type of experiment.

In some experiments, the protocol is intentionally varied as a way of obtaining different realizations. The protocol may be defined in terms of settings of measurement system parameters, and these settings may be varied in the experiment. This way of thinking underlies *ruggedness testing* [2,9,17] and *parameter design* [18]. In addition, the background conditions un-

*Corresponding author. Tel.: +1 301 975 2851; Fax: +1 301 990 4127; E-mail: walter.liggett@nist.gov.

ISSN 0278-0240/04/$17.00

der which the protocol is executed may be varied in the experiment.

Another way of thinking about realizations is in terms of sources of variation. Although some sources of variation cannot be controlled, others can be held constant as successive measurements are made or, alternatively, allowed to vary. The source itself determines the amount of variation. If the controllable sources of variation are held constant, the closeness of agreement between the results of successive measurements is called the *repeatability* [8]. If one or more controllable sources of variation are instead allowed to have their typical effects on the measurement, then the closeness of agreement between successive measurements is called the *reproducibility* [8]. The reproducibility depends on which source or sources are allowed to vary.

Contrasting these two cases constitutes what is called a *gauge R&R* (repeatability and reproducibility) *study* [12]. If the reproducibility is appreciably worse than the repeatability, then an effort should be made to control the relevant sources of variation, usually by changing the measurement protocol. Although other types of experiments are also important, we detail the gauge R & R type of experiment here.

The system development experiment we consider involves measurement of identical subsamples of a human serum standard. The standard is a pooled serum obtained from 250 males and 250 females, all young and healthy. In our experiment, the controllable source of variation is the actualization of a commercially available protein biochip, the IMAC++ metal affinity chip (IMAC stands for immobilized metal affinity capture). Such a chip plays an important role in SELDI-TOF mass spectrometry [11]. The surface of this chip has chromatographic properties that pre-select subsets of serum components. This focuses the analytical readout of the SELDI-TOF mass spectrometer instrument, and thus the spectra produced are cleaner. The chip surface is designed to bind only a subset of proteins from a complex mixture based on physical features such as hydrophobicity, pH or affinity to metal ions. The chip-to-chip variation is that which remains even under ostensibly identical conditions of biochip manufacture and use. The experimental results in this study are groups of measurements. Within each measurement group, the biochip is held constant. Among measurement groups, components of the universe of serum proteins are captured with biochips of differing "bait" surface characteristics. From such results, we can judge how much biochip variation contributes to the measurement variation. If this additional variation were large, we would seek to improve the protocol by reducing the variability in the biochips used and the variability in the process of applying samples to the biochips.

Gauge R&R studies for univariate measurements are routine. Closeness of agreement is measured by the familiar sample variance. In the case of analytical instrumentation such as SELDI-TOF mass spectrometers, we would have a one-dimensional measurement if a feature of the spectrum were identified as the clinically useful biomarker. The feature could be, for example, the intensity corresponding to a particular serum protein. Lacking a feature or even a group of features on which to focus, achieving the goals of gauge R&R studies is more difficult. In this paper, we show how to achieve the goals of gauge R&R without prior information on which features of the mass spectrum are useful as a biomarker. Our approach is based on functional principal components analysis (PCA) [15]. The idea of using PCA in gauge R&R studies appears in Liggett et al. [10]. In this paper, this idea is executed with more refined data analysis and developed more completely.

Lee et al. [9] apply the usual multivariate PCA to HPLC peptide maps to achieve objectives somewhat different from ours. Because functional PCA is similar to multivariate PCA, one can obtain from Lee, et al. [9] different ideas on what can be done with PCA. The application in Lee et al. [9] is ruggedness testing rather than gauge R&R. Their application of PCA emphasizes just dimensionality reduction whereas we also consider interpretation of the individual principal component weight functions as they change with the mass-to-charge ratio. This interpretation requires careful registration of the spectra, a pretreatment step that Lee et al. [9] mention but do not actually perform.

Just as we discuss measurement system development that does not depend on protein information, Petricoin et al. [14] and Adam, et al. [1] undertake demonstration of biomarker feasibility prior to characterization of links between proteins and target disease. Petricoin et al. [14] and Adam et al. [1] analyze collections of case and control specimens with SELDI-TOF mass spectrometry. They use the spectra obtained as training data to derive classifiers that predict the disease state from the observed mass spectrum. The idea of deriving a biomarker based on the mass spectrum without identifying the proteins is attractive. Moreover, Petricoin, et al. [14] and Adam et al. [1] claim that the classifiers they derive perform well. However, Baggerly et al. [3] have raised questions about the classifiers of Petricoin et al. [14].

Billheimer [5] applies functional analysis of variance [15] to MALDI-TOF mass spectra in a case-

control study. He also proceeds without identifying proteins. He shows that the parts of the spectrum most useful in distinguishing cases from controls do not correspond to the obvious spectral peaks and therefore to easily identified proteins.

The classifiers derived by Petricoin et al. [14] and Adam, et al. [1] do contain derived spectral features, and one might imagine use of these for gauge R&R studies. These features, however, depend on the instrument operating conditions during analysis of the case and control specimens. Moreover, these features are somewhat noisy due to the relatively limited number of case and control specimens. Thus, an approach to gauge R&R for spectra that does not depend on case and control specimens is both desirable and useful.

Here we describe gauge R&R analysis that is suitable for a system with functional response, such as a SELDI-TOF mass spectrum analysis. As background, we discuss gauge R&R for an elementary univariate response in Section 2. In Section 3, we discuss the functional case. In Section 4, we address the differences between measurement system studies and case-control studies. The case-control framework is commonly adopted for biomarker development but is quite different from the framework for measurement system studies. We recommend that both be considered in maximizing efforts toward successful biomarker development.

2. Univariate observations

In this analysis, there are 88 measurements of the same human serum standard all made on the same day. These measurements correspond to 8 specimen spots on each of 11 biochips. Before further adjustment as discussed in Section 3, the 88 measurements are mass spectra with baseline correction and normalization as provided by the instrument manufacturer. The average of these 88 mass spectra for a particular mass-to-charge (m/z) interval is shown in Fig. 1. The large peaks dominate perception of this average, but these peaks although prominent, may be of limited use as biomarkers. In terms of the presence of large peaks, the m/z interval depicted in Fig. 1 seems more interesting than the intervals above it or below it. As indicated in Fig. 1, we have chosen only m/z in the subinterval [7690, 9389] for consideration in Section 3. In this section, we consider just the largest peak, the one at $\text{m/z} = 7775$ on the right side of the interval chosen for Section 3.

We begin our investigation of our 88 mass spectra with a rudimentary analysis. We quantitate the largest of the spectral peaks and treat the resulting univariate measurements as data of a gauge R&R study. As well as being an illustration, this would actually be appropriate to biomarker development if the largest peak corresponded to a serum protein concentration directly reflective of a biological characteristic such as the presence of early cancer. Reduction of the spectra to the values of several peaks is a common first step in data analysis for chemical spectra. Univariate gauge R&R is widely used for measurement system development. Our rudimentary analysis reveals data properties that help justify steps we take when working with the mass spectra in the next section.

We extract from each of the 88 spectra the height of the largest peak. We find the height by interpolating the spectral points in the vicinity of the peak with a cubic spline and finding the maximum of the interpolating function. There are certainly other ways of quantitating the peaks of chemical spectra such as finding the height from a baseline or finding the peak area. However, for the purpose of illustrating univariate gauge R&R, the method we have chosen is suitable.

Figure 2 shows the height of the largest peak for each measurement. Each letter on the vertical axis of this dot plot corresponds to an independent biochip. Thus, there are 11 letters each with 8 dots corresponding to the spots on the biochip. We note that biochip F includes 3 spots with high values. Further investigation shows that the corresponding spectra have much higher baselines than the other spectra. As part of our analysis in Section 3, we correct for this. Quantifying the spread of values for each biochip by computing the standard deviation and pooling the 11 results provides an estimate of the repeatability. Figure 2 shows that between biochips there is some extra variability. For example, the values for biochip B are all lower than the values for biochip H. This indicates that biochip as a reproducibility factor, adds variability. An estimate of the standard deviation that would be observed if this factor were allowed to vary generally in a sequence of measurements can be computed from the values in Fig. 2. Thus, the increase attributable to biochip variation can be computed [12].

3. Functional observations

Extension of the analysis in Section 2 to the 7690–9389 m/z subinterval indicated in Fig. 1 is important when it is unknown what features of the SELDI-TOF

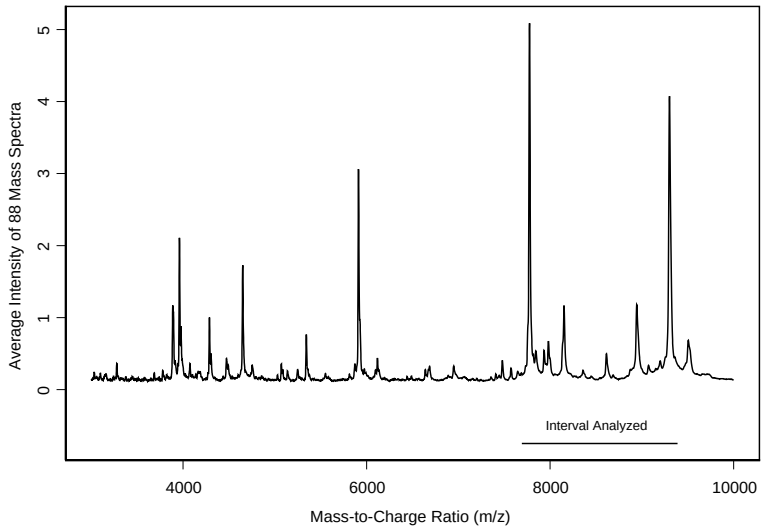

Fig. 1. Average spectrum for an m/z interval with prominent peaks.

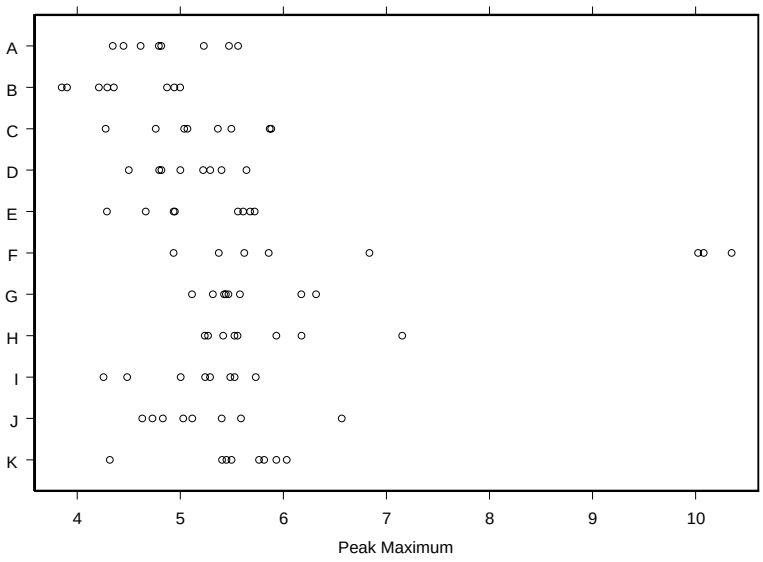

Fig. 2. Plot of heights of largest peak grouped by biochip.

mass spectrum are useful as biomarkers. For this subinterval, Fig. 1 shows a few major peaks and several minor ones, any of which might be important. Thus, we would like to perform gauge R&R in a way that takes into account the entire curve. In other words, we would like to see all the effects that the reproducibility factor has on the functional measurement. This requires replacing the standard deviation used as an indicator of variability in the univariate case with something else that takes into account the entire curve. Functional PCA has potential as an indicator of curve variability [15]. Of course, functional measurements are inherently more complex than univariate measurements, and functional PCA will not eliminate all the additional complexity. In particular, only under exceptional circumstances can curve variability be described by a value analogous to a single standard deviation.

Functional PCA provides insight by summarizing a batch of functional measurements in terms of typical deviations from the mean of the batch. These typical deviations are the principal component weight functions. This is somewhat like summarizing a batch of univariate measurements by the mean and standard deviation except that both the mean and the deviations are

curves. Functional PCA provides an indication of how many alternative forms of deviation are represented in the batch. For each form of deviation, that is, each principal component weight function, there is an indication of its strength at each value of the independent variable. In place of covariance functions, functional PCA provides an accessible interpretation of the structure of measurement-to-measurement variability. In mathematical terms, it provides an informative low-dimensional representation of the batch.

Typically, functional measurements are presented for analysis in sampled form, that is, as observations at discrete values of the independent variable u. For our data, the variable u denotes mass-to-charge ratio. We do not use t as the symbol for the independent variable as Ramsay and Silverman [15] do because t might be associated with time, which is an alternative independent variable for time-of-flight mass spectra. There are N functional measurements. Measurement i is presented as n values $y_{i1}, \ldots, y_{in}$, where y_{ij} is the observation of the function i at u_j. If as we assume, the sampling is fine enough and the measured function is at least continuous, we can think of $y_{i1}, \ldots, y_{in}$ as a function y_i with values $y_i(u)$ computable for any value of the independent variable u, and use these functions in the calculations for functional PCA. See Ramsay and Silverman [15] for the necessary software. A cubic-spline representation of the function y_i derived from $y_{i1}, \ldots, y_{in}$ is the basis of the data analysis in this paper.

Note that we could apply principal components analysis directly to the matrix with elements y_{ij}. However, in the case considered here, we must treat the data functionally so that we can interpolate between the given spectral values to register the curves. Also, treating the data functionally allows smoothing the spectra or alternatively, applying functional PCA in a way that produces smooth weight functions [15]. We have not done this here, but it is a potentially important option.

3.1. Adjustment of the spectra

To make the principal components interpretable, we need to remove from the spectra some of the variation. Consider Fig. 3(a), which shows the 88 spectra in the vicinity of the peak discussed in Section 2. Three spectra stand out as having a problem with their baselines as mentioned in connection with Fig. 2. We adjust the baselines for all the spectra including these three. Also, the location of the peak does not occur at the same m/z value for each spectrum. We register the spectra to remove this source of variation. Although it cannot be identified in Fig. 3(a), there may be measurement-to-measurement variation in the amount of protein released from the chip by the laser. We normalize the spectra to adjust for this. Finally, in the course of an initial analysis of these measurements, we identified an outlier. The outlying measurement was produced from spot 8 on chip J. Although this outlying measurement may be interesting, treating it separately is better than including it in the PCA. Thus, we have $N = 87$.

Because the 7690–9389 m/z interval is relatively small, we adjust the baseline by subtracting from each spectrum the average of the spectrum over the interval. The baseline-adjusted spectra are given by

$$z_i(u) = y_i(u) - \int_{u_1}^{u_2} y_i(s)ds/(u_2 - u_1),$$

where $u_1 = 7960$ and $u_2 = 9389$.

Registration of a spectrum can be thought of as replacement of $z_i(u)$ with $z_i(h_i(u))$. As a form for the registration, we take $h_i(u) = \delta_i + \gamma_i u$. The registration affects the baseline adjustment in that

$$z_i(h_i(u)) = y_i(h_i(u)) - \int_{u_1}^{u_2} y_i(h_i(s))ds/(u_2 - u_1).$$

Note that the registered spectra may include values somewhat outside the 7690–9389 m/z interval. This is not a problem because the mass spectra extend beyond the interval we have chosen.

Registering spectrum i consists of estimating δ_i and γ_i. As estimates, we follow Ramsay and Silverman [16] and find the values of δ_i and γ_i that minimize the smaller of the two eigenvalues of the matrix

$$\begin{bmatrix} \int_{u_1}^{u_2} \{z_0(s)\}^2 ds & \int_{u_1}^{u_2} z_0(s) z_i(h_i(s)) ds \\ \int_{u_1}^{u_2} z_0(s) z_i(h_i(s)) ds & \int_{u_1}^{u_2} \{z_i(h_i(s))\}^2 ds \end{bmatrix}.$$

The rationale for this estimation criterion stems from the fact that if $z_i(h_i(u)) = Az_0(u)$, then the smallest eigenvalue is 0. In other words, this criterion seeks to register spectrum i so that it is proportional to the model spectrum $z_0(u)$.

We determine $z_0(u)$ by what is called the Procrustes method [15]. Initially, we take

$$z_0(u) = \frac{1}{N} \sum_{i=1}^{N} z_i(u).$$

Then, after registration on the basis of this value of $z_0(u)$, we replace it with

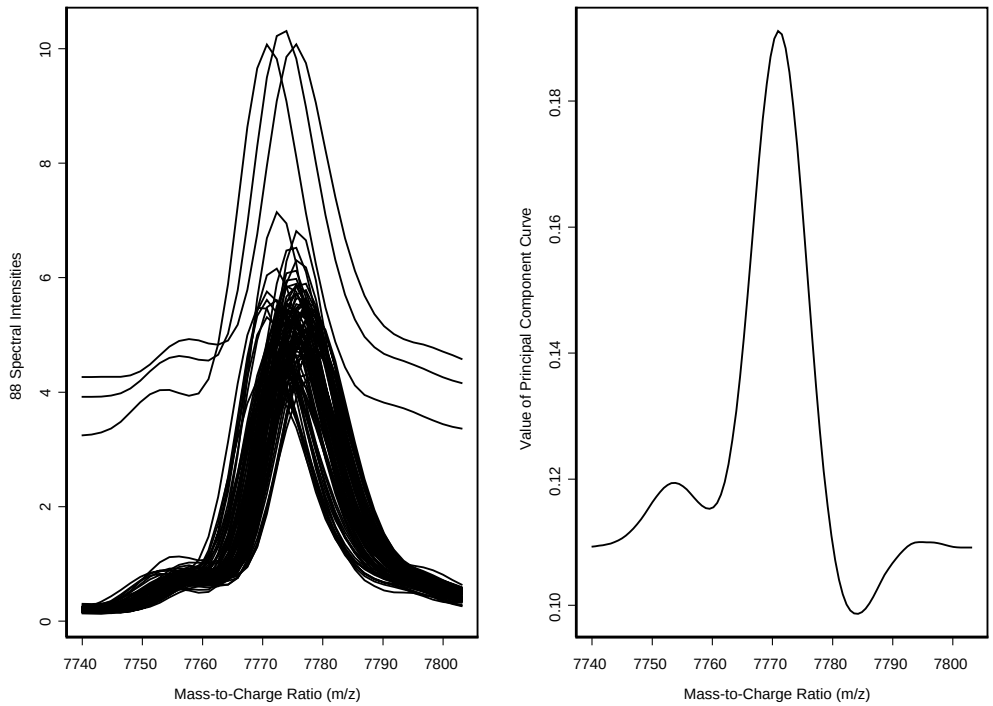

Fig. 3. Spectral curves for an interval containing the largest peak and the corresponding principal component weight function.

$$z_0(u) = \frac{1}{N}\sum_{i=1}^{N} z_i(h_i(u))$$

and refine the estimate of the registration. Note that in computing $z_0(u)$, we omit the outlier.

We continue this until refining the registration seems to make little difference. The final value of $z_0(u)$ we denote by $\overline{z}(u)$.

The function $z_i(h_i(u))$ is spectrum i after baseline correction and registration. Now we want to account for the possibility that the amount of material desorbed from the chip changes from measurement to measurement. The reason is that we do not want to have to distinguish this source of variability from what else PCA has to show. We model the effect on a spectrum of variability in material desorbed as addition or subtraction of an amount proportional to $\overline{z}(u)$. As a consequence of this model, we adjust for this source of variability by computing

$$x_i(u) = z_i(h_i(u)) - \overline{z}(u)\left[\frac{\int_{u_1}^{u_2} z_i(h_i(s))\overline{z}(s)ds}{\int_{u_1}^{u_2}\{\overline{z}(s)\}^2 ds}\right].$$

Note that after this adjustment, spectrum i is given by $x_i(u)+\overline{z}(u)$, that is, the $x_i(u)$ to which we apply PCA are deviations from the mean spectrum. The advantages of an additive adjustment become apparent when we want to interpret the principal component weight functions.

The spectra after adjustment $(x_i(u) + \overline{z}(u))$ are shown superimposed in Fig. 4. This figure provides a rough picture of the variation as it compares to the average of the spectra. The amount of variation as shown in this figure must be considered in the interpretation of subsequent figures where the variation is exaggerated so that details can be seen. We note that the m/z values where the variation is larger correspond to the spectral peaks. However, the amount of variation is not proportional to the mean $\overline{z}(u)$. The smaller peak at m/z $= 8944$ has as much variation as the large peaks at m/z $= 7775$ and m/z $= 9299$.

3.2. Functional PCA

As an introduction to PCA, consider the following: Functional measurements such as chemical spectra are usually analyzed by first reducing each measurement to one or more features. One type of feature is a linear combination of the function values for different values of u. A feature of this type is given by

$$f_i = \int \beta(s)[x_i(s) + \overline{z}(s)]ds, \quad i = 1, \ldots, N.$$

The weight function β defines the feature f. One could specify β and analyze the resulting batch of numbers $f_i, i = 1, \ldots, N$. Note that by proper selection of β, one could obtain areas under a peak as the values f_i.

Rather than specifying β in advance, one can ask for the weight function that gives the feature most descriptive of the variation of the functional measurements. More explicitly, one can find the weight function ξ_1 for which the corresponding feature values

$$f_{i1} = \int \xi_1(s) x_i(s) ds$$

have the largest possible $\sum_i f_{i1}^2$ subject to $\int \xi_1^2(s) ds = 1$. This gives the weight function ξ_1 and the scores $f_{i1}, i = 1, \ldots, N$ for the first functional principal component.

Returning to Fig. 3, we note that even without adjusting the 88 spectra shown in Fig. 3(a), we can center this collection of curves and find the first principal component for the centered curves. The weight function of the first principal component ξ_1 is shown in Fig. 3(b). We see that ξ_1 heavily weights the peak, which as shown in Fig. 3(a) is the most variable part of this batch of functional measurements. Such a relation between the peaks and the principal components is obvious in our application of PCA to the $x_i(u), i = 1, \ldots, N$.

A more complete description of the variation in a batch of functional measurements can be obtained by computing more weight functions. Proceeding step by step, we compute at step k, the weight function ξ_k corresponding to feature values

$$f_{ik} = \int \xi_k(s) x_i(s) ds.$$

The weight function ξ_k is chosen to maximize $\sum_i f_{ik}^2$ subject to $\int \xi_k^2(s) ds = 1$ and the additional $k - 1$ constraints

$$\int \xi_m(s) \xi_k(s) ds = 0$$

for $m < k$. We see that functional PCA extracts features of the variability in descending order of sample variance and that the weight functions are uncorrelated.

The idea that the principal components describe the variation of the batch can be made clearer. For any fixed K, let

$$\hat{x}_i(u) = \sum_{k=1}^{K} f_{ik} \xi_k(u) \tag{1}$$

be an expansion of x_i in terms of the first K principal component weight functions ξ_k. In terms of integrated square error

$$\sum_{i=1}^{N} \int [x_i(s) - \hat{x}_i(s)]^2 ds,$$

this expansion gives at least as good an approximation to the batch of functional measurements as any other set of K functions. We can quantify the amount of variation explained by the first principal K components by means of the quantity

$$1 - \left\{ \sum_{i=1}^{N} \int [x_i(s) - \hat{x}(s)]^2 ds \Big/ \sum_{i=1}^{N} \int [x_i(s)]^2 ds \right\}.$$

3.3. PCA results

We now apply functional PCA to the $x_i(u)$. The first and second principal components ξ_1 and ξ_2 and the mean $\overline{z}$ are shown in Fig. 5. These two principal components account for 70% of the variation in the spectra. With respect to the amount of variation actually observed in the spectra, the scale of these principal components is exaggerated relative to the scale of the mean.

As shown in Fig. 5, the first principal component has peaks that correspond to the spectral peaks shown by the mean spectrum. Billheimer [5] made a similar observation about his data. As discussed further in the Appendix, that the peaks in this principal component are positive and negative is in part the result of the spectral adjustment. As part of the adjustment, we have forced the principal components to be orthogonal to a constant and to $\overline{z}$. Note that the magnitudes of the peaks in this principal component are not proportional to the magnitudes of the peaks in the mean. This observation corresponds to a similar observation about Fig. 4.

The second principal component seems to reflect more than increased variability at the spectral peaks. The second principal component swings rapidly back and forth with m/z in the vicinity of the two largest spectral peaks. This behavior is what one might expect if the spectral variation is in part the broadening and narrowing of these two peaks. Moreover, the second principal component seems to be associated with variation in the smaller spectral peaks between the two large ones. Further interpretation of the principal components is found in the Appendix.

Let us return to the goal of understanding how much the variation from biochip to biochip contributes to the variation in the spectra. For each principal component k, each (centered) spectrum i is associated with a score f_{ik} that gives the amount of the spectrum that can be at-

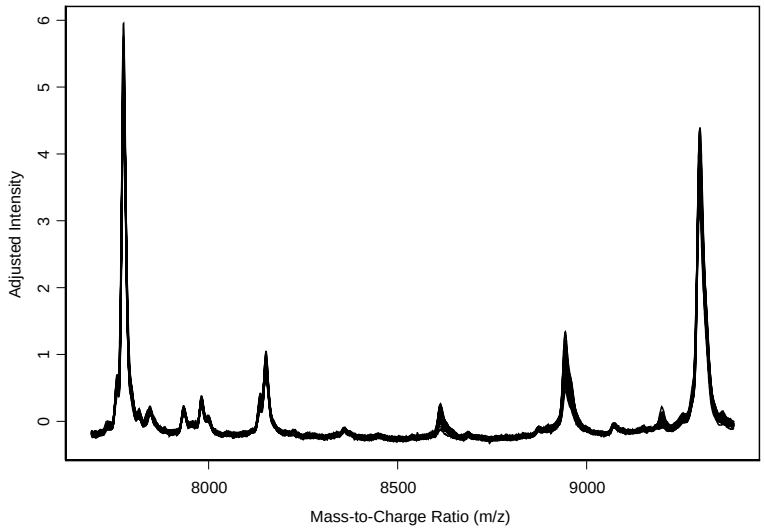

Fig. 4. Eighty-seven spectra registered, centered, and normalized.

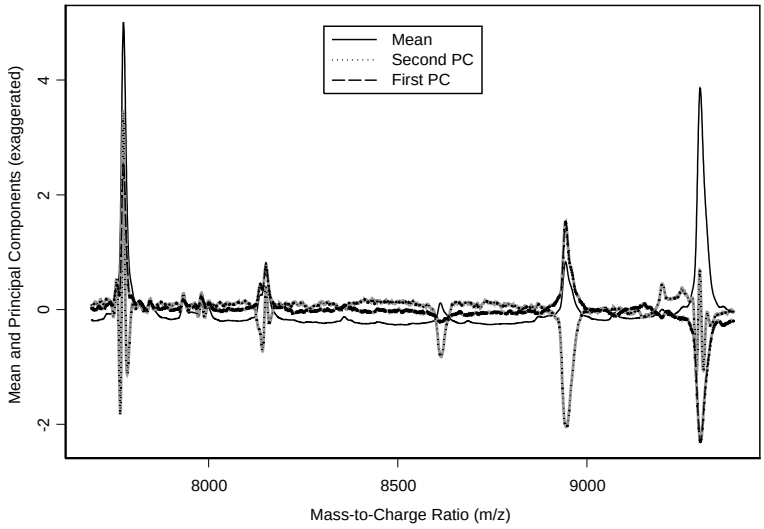

Fig. 5. First and second principal components compared to the mean spectrum.

tributed to the principal component. Plotting the scores for the first principal component versus the scores for the second is a common data analysis step [15]. If there were no association between the locations of the points on this plot and the biochip from which the spectrum came, then we would conclude, at least to the extent of the first two principal components, that there is no additional variability contributed by variation in the biochip. Plotting scores for other pairs of principal components might also be informative. Figure 6 shows such a plot with the points labeled by biochip. Apparently, there is some association. Moreover, compared to the first principal component, the dependence on biochip seems greater for the second principal component. For example, in terms of the second principal component, biochip H has high scores and biochip B has low scores. This means that were we to compare the spectra from biochip H with the spectra from biochip B, we would see variation similar to what is shown in exaggerated form in Fig. 5 for the second principal component. If we could identify the sources of variation reflected in the second principal component, then we might hypothesize the mechanism behind the chip-to-chip differences and possible remedies.

Figure 6 provides a way of comparing the spot-to-spot variability for a given chip with the chip-to-chip

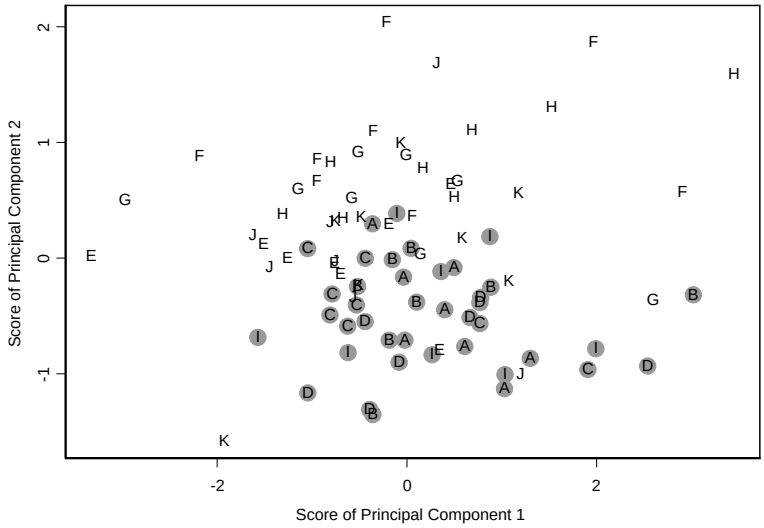

Fig. 6. Scores for first and second principal components with biochip indicated.

variability. In this regard, Fig. 6 provides for functional measurements what Fig. 2 provides for univariate measurements. In terms of the definitions of repeatability and reproducibility, Fig. 6 shows "closeness of agreement" but in a way that depends on the principal component weight functions, which are derived from the data. One can ask whether the quantification of variability possible in the univariate case is also possible in the functional case. The problem with such quantification in the functional case is lack of knowledge of the relation between the principal component weight functions and the biomarker, which is as yet undiscovered or at least not completely specified. Thus, maybe all we say is that Fig. 6 shows chip-to-chip variability the reduction of which is generally a good idea.

We have, of course, analyzed just part of each spectrum. For the particular standard used, most of the prominent peaks are in the m/z interval shown in Fig. 1, and we have examined only a fraction of this interval. Consider applying PCA to a larger part of the interval shown in Fig. 1. In this case, interpretation would involve more principal components because there would be more peaks. For this reason and because figures such as Fig. 5 would become unreadable, interpretation would be more difficult. Consider applying PCA to an interval that extends above the interval shown in Fig. 1. In this extension, the intensities are much lower but of interest. In this case, the large peaks shown in Fig. 1 would dominate the results of PCA. For this reason, interpretation of the spectra in the extension would be difficult. Consider applying PCA to successive nonoverlapping intervals. This might be better than application of PCA to a larger interval, but correlation between spectral features in different intervals would not enter the analysis. For this reason, some important aspects of the spectra might be missed. Clearly, there is a need to expand the methodology described in this paper so that larger sections of the spectra can be better analyzed.

4. Metrological and case-control studies

The foregoing description of measurement protocol improvement differs from the case-control studies that have been put forward as a paradigm for biomarker development [13]. In a case-control study, one obtains a group of samples from individuals with the disease in question and a group of samples from individuals free of the disease in question. One then observes the category into which the biomarker places the individuals. A biomarker with high sensitivity is one with a high rate of true positives, and a biomarker with high specificity is one with a low rate of false positives. Clinical validation of a biomarker involves case-control studies. The experiment described in Section 3 is designed for analytical validation and does not address clinical validation.

The difference between metrological studies and case-control studies can be characterized in three ways. First, in metrological studies, one generally measures the same specimen twice whereas in case-control studies, one contrasts measurements of specimens from diseased individuals with measurements of specimens

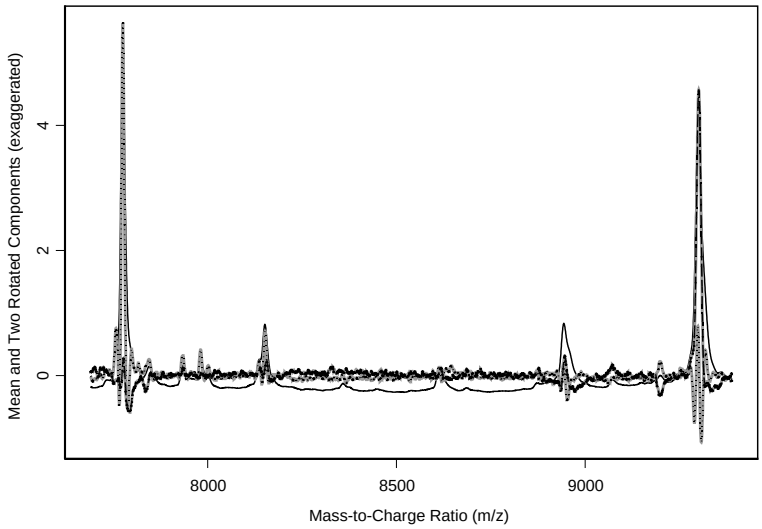

Fig. 7. Rotated components associated with left and right peaks compared to mean spectrum.

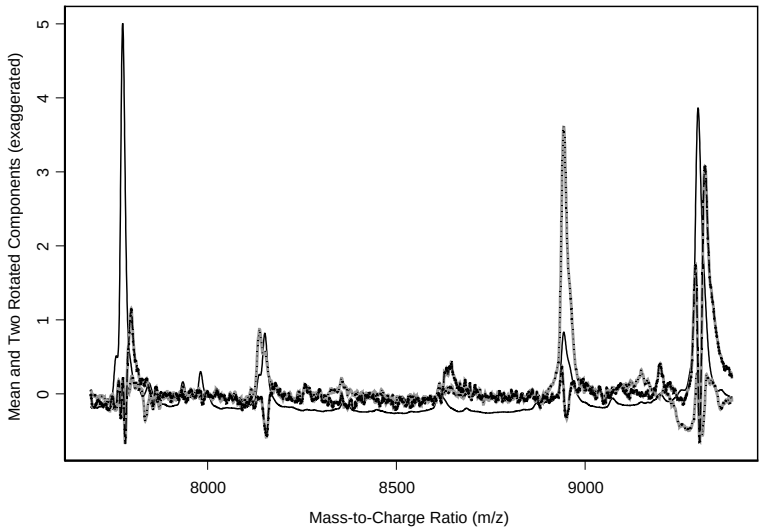

Fig. 8. Rotated components associated with middle peaks and trailing edges of peaks compared to the mean spectrum.

from healthy individuals. Second, the difference stands out in terms of the scientific basis. Measurement protocol evaluations have their basis in scientifically defined quantities such as protein concentrations. Case-control studies have their basis in a scientifically sound way of distinguishing diseased individuals from healthy individuals. Third, specimens for case-control studies are more expensive to procure because they must properly reflect diseased and disease-free populations [13]. Many metrological studies, on the other hand, require little but specimens that can be divided into homogeneous subsamples.

Consider the development of a biomarker based on measurement of human serum. Say that we have serum specimens known to be from diseased individuals and other specimens known to be from healthy individuals. One can proceed with the development as Petricoin et al. [14] and Adam, et al. [1] did. One can measure as many properties of each specimen as one would like. The measurements on a specimen could be a SELDI-TOF mass spectrum as in Petricoin et al. [14] and Adam et al. [1]. One could then search for the combination of these measurements that best distinguishes cases from controls. If one shows that this combination is indeed effective, then it can be used as a biomarker. The ef-

fectiveness of such an approach, however, depends on the quality of the measurements. In fact, with poor measurements, such an approach might not reveal any promise of a new biomarker. In contrast, Baggerly, et al. [3] discuss measurement problems perhaps leading to a false indication of biomarker effectiveness. Thus, experiments to improve measurement quality should not be postponed. Early in the development, one might try to find some indication that the instrumentation in some configuration is capable of distinguishing cases from controls. Beyond that, however, one would not want to take the instrument configuration, and more broadly the measurement protocol, as fixed. Measurement protocols as initially specified are based in part on engineering judgments. Considerable improvement may result from checking these judgments experimentally. *The best advice on the steps to be undertaken in an engineering project stipulates that the first step be study of the measurement system [7]*. The approach described in this paper allows study of the measurement system before the interpretation of the response is finalized.

Even after an effective biomarker has been discovered, more is necessary in biomarker development. One must be concerned with the variability associated with the measurement protocol when implemented by several laboratories as the use of the biomarker becomes wide spread. This concern should be addressed through the sorts of experiments that metrologists perform. One might take a set of case-control specimens, split each of them, and measure the two identical sets with different protocols to see which gives the best sensitivity and specificity. If one spreads the measurement of each set among several laboratories, one might see which protocol was best in an interlaboratory setting. However, this approach seems cumbersome at best. Moreover, it makes no use of knowledge of the measurement mechanism. Generally, this part of biomarker development should more closely follow metrological practice.

In the instance of human-serum biomarkers, a scientific description of the proteins that form the basis of the biomarker would be ideal. This should lead to a scientific understanding of how the biomarker is connected with the condition it is intended to detect and thereby improvement of the biomarker. This might also lead to focus on specific aspects of the measurement mechanism and thereby more effective metrological experiments. But knowledge of the specific proteins is not necessary to achieve effective class discrimination nor is it an essential component of a biomarker. Despite questions about their specific performance findings, Petricoin et al. [14] and Adam, et al. [1] have shown how a case-control study can proceed without this knowledge. In Section 3, we demonstrate how metrological studies can proceed without this knowledge.

In structuring biomarker development, one must be sure that the distinct difference between case-control studies and metrological studies does not become an impediment to planning. Because of the difference, integrating both study types is challenging. Each has the potential for contributing to the biomarker development, and reducing the time from discovery to clinical application. Moreover, their relative timing in biomarker development sequence is an important issue. Because the required components of each approach are so distinct, it may be difficult to derive a commonly accepted plan for successfully combining case-control and metrological studies. A general sequence of phases such as those advocated by Pepe et al. [13] may not be amenable to achieving the separate objectives. However, this lack of a consensus, if recognized, can be overcome so that it does not slow biomarker development.

Acknowledgement

We are grateful to Dr. David L. Duewer of NIST for critical comments and suggestions.

Disclaimer

Certain commercial entities, equipment, or materials may be identified in this paper in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Appendix. Interpretation of Principal Components

PCA would lead to greater understanding if each component could be associated with a mechanism responsible for spectral variation. The first and second principal components, which are shown in Fig. 5, seem to be the result of several mechanisms and therefore hard to interpret. It is possible to transform the first

K principal components into a new set of components that are more interpretable in terms of mechanisms. The transformation is called VARIMAX rotation [15]. This appendix presents the results of this transformation. Among other things, these results suggest the mechanisms that influence the gauge R&R plot shown in Fig. 6.

VARIMAX rotation could be applied directly to the first K principal components, but supplementing this set can lead to better results. Because of the centering and normalization of the spectra, the principal components are orthogonal to the constant spectrum and to the mean spectrum. Let $\xi_{K+1}(u)$ be a constant that is determined by $\int \xi_{K+1}^2(s)ds = 1$, and let $\xi_{K+2}(u)$ be proportional to the mean spectrum $\overline{z}(u)$ with the constant of proportionality given by $\int \xi_{K+2}^2(s)ds = 1$. We can rewrite Eq. (1) with these two extra components

$$\hat{x}_i(u) = \sum_{k=1}^{K+2} f_{ik}\xi_k(u)$$

by letting $f_{ik} = 0$ for $k = K + 1$ and $k = K + 2$. Applying VARIMAX rotation to this representation gives

$$\hat{x}_i(u) = \sum_{k=1}^{K+2} g_{ik}\psi_k(u),$$

where the $\psi_k(u)$ are the rotated components. The reason for including the two extra components is that VARIMAX rotation attempts to find rotated components that are large over a small portion of the interval and nearly zero elsewhere. Thus, for example, one would hope that VARIMAX rotation would lead to a separate component for each of the major spectral peaks. With the two extra components, the VARIMAX rotation is not constrained by orthogonality to the constant spectrum and the mean spectrum. Thus, the goal of VARIMAX rotation can be more nearly achieved. Note that the rotated components $\psi_k(u)$ are not orthogonal to the constant spectrum and the mean spectrum as the original principal components are.

In computing the rotated components, we set $K = 8$. We add the extra two components and rotate the result. In Figs 7 and 8, we show four of the resulting ten components. The other components contribute little to the representation of the variation of the spectra. Note that because functional PCA followed by VARIMAX rotation is data driven, the rotated components that result will not perfectly match what might be expected from an understanding of the measurement system. Thus, we cannot be sure of our interpretation.

Figure 7 shows two components, one that depicts the variation associated with the leftmost peak and another that depicts the variation associated with the rightmost peak. Figure 8 shows two components, one that depicts the variation associated with two of the middle peaks (dotted) and one that depicts variation associated with the trailing edge of the peaks in the interval (dashed). That one component in Fig. 8 shows variation associated with peaks at m/z = 8152 and m/z = 8944 suggests correlation between the variations of these peaks. In other words, if the two peaks were uncorrelated, one would expect them to rotate into separate components. If such correlation were large, it would be of interest. We conclude that the spectral variation shown crudely in Fig. 4 is largely associated with mechanisms that influence the heights of the peaks and mechanisms that influence the widths of the peaks. One could do a more detailed comparison of the four weight functions shown with known mechanisms that contribute to the shapes of SELDI-TOF mass spectra.

References

[1] B.L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Y. Yasui, Z. Feng and G.L. Wright, Jr., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperlasia and healthy men, *Cancer Research* **62** (2002), 3609–3614.

[2] K.D. Altria, B.J. Clark, S.D. Filbey, M.A. Kelly and D.R. Rudd, Application of chemometric experimental designs in capillary electrophoresis: A review, *Electrophoresis* **16** (1995), 2143–2148.

[3] K.A. Baggerly, J.S. Morris and K.R. Coombes, Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments, *Bioinformatics* **20** (2004), 777–785.

[4] P.E. Barker, Cancer biomarker validation: Standards and process, *Annals of the N. Y. Academy of Sciences* **983** (2003), 1–9.

[5] D. Billheimer, *A functional data approach to MALDI-TOF MS protein analysis*, Poster presentation at the University of Florida Fifth Annual Winter Workshop, January 10–11, 2003, http://web.stat.ufl.edu/symposium/2003/fundat/Archive.

[6] Biomarkers Definitions Working Group, Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework, *Clinical Pharmacology and Therapeutics* **69** (2001), 89–95.

[7] G.J. Hahn, W.J. Hill, R.W. Hoerl and S.A. Zingraf, The impact of six sigma improvement-A glimpse in the future of statistics, *The American Statistician* **53** (1999), 208–215.

[8] International Organization for Standardization (ISO), *International Vocabulary of Basic and General Terms in Metrology*, (2nd ed.), Geneva, Switzerland, 1993.

[9] K.R. Lee, J. Bongers, B.H. Jones and S. Burman, Ruggedness study of HPLC peptide mapping for the identity of a drug compound: A chemometrics approach, *Drug Development and Industrial Pharmacy* **26** (2000), 123–134.

[10] W. Liggett, L. Cazares and O.J. Semmes, A look at mass spectral measurement, *Chance* **16**(4) (2003), 24–28.
[11] M. Merchant and S.R. Weinberger, Recent advances in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry, *Electrophoresis* **21** (2000), 1164–1177.
[12] D.C. Montgomery, *Introduction to Statistical Quality Control*, (2nd ed.), John Wiley, New York, 1991.
[13] M.S. Pepe, R. Etzioni, Z. Feng, J.D. Potter, M.L. Thompson, M. Thornquist, M. Winget and Y. Yasui, Phases of biomarker development for early detection of cancer, *Journal of the National Cancer Institute* **93** (2001), 1054–1061.
[14] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn and L.A. Liotta, Use of proteomic patterns in serum to identify ovarian cancer, *The Lancet* **359** (2002), 572–577.
[15] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, Springer-Verlag, New York, 1997.
[16] J.O. Ramsay and B.W. Silverman, *Applied Functional Data Analysis: Methods and Case Studies*, Springer-Verlag, New York, 2002.
[17] Y. Vander Heyden, C. Hartmann, D.L. Massart, L. Michel, P. Kiechle and F. Erni, Ruggedness tests for a high-performance liquid chromatographic assay: comparison of an evaluation at two and three levels by using two-level Plackett-Burman designs, *Analytica Chimica Acta* **316** (1995), 15–26.
[18] C.F.J. Wu and M. Hamada, *Experiments: Planning, Analysis, and Parameter Design Optimization*, John Wiley, New York, 2000.