ELSEVIER

Intern. J. of Research in Marketing 19 (2002) 167–179

International Journal of
Research in
Marketing

www.elsevier.com/locate/ijresmar

Exploring repeated measures data sets for key features using Principal Components Analysis

Eric T. Bradlow*

Department of Marketing and Statistics, The Wharton School of the University of Pennsylvania, Suite 1400 SH-DH, 3620 Locust Walk, Philadelphia, PA 19104, USA

Received 10 October 2001; received in revised form 1 February 2002; accepted 8 February 2002

Abstract

Repeated measures data sets are a very common data structure in marketing, ranging from weekly household purchase data obtained via supermarket scanners, household penetration rates for products over time, panel survey questionnaire responses, etc. In many circumstances, it is desirable to glean key features of the data without having to run a formal model and/or perform a large number of exploratory analyses. Such features of interest may include defining market segments, identifying change points (a sudden shift in aggregate behavior), and, in general, identifying "typical" behavior patterns. In this research, we apply an exploratory approach for analyzing repeated measures data sets, initially considered by Tucker [Psychometrika 23 (1958) 19] and fully explicated in Jones and Rice [Am. Stat. 46 (1992)], that utilizes Principal Components Analysis (as its core), regression, and simple bivariate plots. It is clearly our intention, via the algorithm provided, that a user can read this manuscript and then sit down at his or her computer and implement the approach immediately. We first demonstrate our approach on simulated data, therefore providing a "blueprint" for users of this approach, i.e. we simulate different types of structural variation and provide the corresponding principal component rotation curves, and how to glean features from them. Our approach is then applied to a data set on cumulative penetration rates for a set of durable products.

Keywords: Exploratory inference; Functional analysis; Standard software

1. Introduction

Of course, managers, when making inferences from data, want to "get it right," but at what cost? Complex statistical models have become omnipresent in the marketing, statistics, and management domains, among others, in many cases to the exclusion of very simple approaches. Imagine a manager interested in detecting homogeneous segments of customers to target. Do we "need" to fit a latent class model? Imagine a manager interested in assessing whether seasonal fluctuations in sales exist and, if so, to what extent? Do we "need" to fit a seasonally adjusted time series model? These questions, among others, have the answer: "it depends." In some cases, "hard-and-fast" answers are available which might suffice for the decision a manager is currently facing. In this research, we focus specifically on data sets composed of repeated measures and an associated "hard-and-fast" approach for answering some basic questions from the data.

* Tel.: 1-215-898-8255.
E-mail address: ebradlow@wharton.upenn.edu (E.T. Bradlow).

0167-8116/02/$ - see front matter
PII: S0167-8116(02)00065-4

Whether it is supermarket scanner data for a sample of households (Russell & Kamakura, 1994) or product penetration rates (diffusion) over time (Van den Bulte, 2000), repeated measures data sets have become ubiquitous in the marketing/management domain. Such databases provide a rich source of information for formal model analysis which, via specialized, easy-to-use software, are implemented by many managers and analysts. However, by their longitudinal nature, such data sets are also very amenable to simple exploratory analyses which, due to the intrigue and "high power" of complex models, are often overlooked. While in most cases exploratory analyses by no means replace the need for more formal models, they can suggest a starting point for more formal models especially with regards to reasonable functional forms. As we describe next, that was the initial motivation behind this approach described in Tucker (1958).

Despite optimism regarding exploratory analyses, the vast number of possible choices can be overwhelming and, unfortunately, does not typically lead to a concise set of inferences; moreover, to be useful requires a carefully thought-out strategy to "navigate" the data. For example, if honest, most analysts will admit to taking repeated measures data, plotting in a single graph a connected symbol chart (Cleveland, 1994, pp. 181–185) with one curve for each unit (hence, we use the term "curve" to refer to a unit's repeated measures) and stating "This is horrible!". The vast mass of curves (ink) makes such an approach mostly useless; however, we all do it in the hope that something will emerge. Thinking carefully, what is obstructing a clear view of the "action" in the curves is that: (a) many curves are really just duplicates of each other measured with error and (b) what the typical curve looks like is hidden. In this research, we provide a general approach, the use of Principal Components Analysis (PCA), followed by simple bivariate plots and ordinary regression, to explain the variation in the curves across units and time. A nice feature of our proposal is that anyone (manager, researcher, student, etc.) can just sit down and "do it" as its analysis components are available in every standard statistical software package and have been run by each of us numerous times. Our approach is an application of the procedure proposed by Jones and Rice (1992) who used PCA to explain the variation in continuous curve data. This idea, however, is attributable to an older source (Tucker, 1958, 1968; Tucker & Messick 1963) that considers the PCA approach as an exploratory procedure for identifying functional forms. A more recent set of research on functional data analysis also exists such as Ramsey and Silverman (1997). Thus, while our research is not methodologically innovative, it does provide a managerially focused application of this procedure, applied to marketing/management issues.

Understanding the sources of variation in repeated measure data sets has much practical use and has been the focus of much current formal modeling work. This includes identifying customer segments (in our parlance, the number of typical curves and what they look like) via latent class methods (Jain, Bass, & Chen, 1990), longitudinal trends across measures, and hierarchical models allowing for heterogeneity across units (Rossi & Allenby, 1993). Although by no means meant as a replacement for any of these formal methods, our approach will be especially useful if the data set is very large, there are time constraints, or if only general inferences are desired. However, on the caveat side, as the PCA approach takes linear combinations of the underlying measures, only features that can be represented this way will be found with any significant power. We discuss this issue in detail.

The remainder of this paper is laid out as follows. In Section 2, we provide a detailed algorithm describing our approach. In Section 3, we simulate data sets with interesting variation structures across units and time and demonstrate how our approach would find such structures. We also include a power study based on an ad hoc selection rule that provides some information regarding the efficacy of the approach. Section 4 applies our methods to real data on a set of diffusion rates for industrial products. A summary and concluding remarks are given in Section 5.

2. The PCA approach

The basic input to our approach is a rectangular data structure $\boldsymbol{Y}$ with no missing data consisting of T measurements for each of I units, i.e. the ith row of $\boldsymbol{Y}$ denoted y_i is the repeated measures data for unit i with measurements $y_{i,t}$. We call y_i the "curve" for unit i. In practice, it may be somewhat restrictive to not allow for ragged data arrays and to require all units to have

exactly T observed measurements. In such situations, one can: (a) only use units with at least T measurements, (b) impute missing data for the unobserved measurements (Little & Rubin, 1987), and/or (c) chop off all measurements after T have been obtained (note, it need not be the first T or even consecutive measurements). We strongly recommend the use of proper missing data methods as (a) or (c) can have severe biasing effects. A step-by-step description of our approach is provided next.

2.1. Step 1: run a PCA on $\boldsymbol{Y}$

Using your favorite software package, run a PCA on $\boldsymbol{Y}$ requesting as output the $T \times T$-dimensional matrix of PCA rotation vectors (denoted by R with columns r_t) and $I \times T$-dimensional matrix of PC scores (denoted by S with columns s_t).

2.2. Step 2: determining the number of important rotated dimensions

For each column s_t, compute the variance $v_t = \mathrm{var}(s_t)$ of the rotated scores on dimension t and their sum $V = \sum_{t=1}^{T} v_t$; of course, $v_1 \geqslant v_2 \geqslant \cdots \geqslant v_T$. Decisions regarding the number of "important" dimensions are usually obtained by keeping all dimensions such that $v_t / V \geqslant c$. Typical choices of c include: (a) $c = 1/T$, keep all dimensions with variance explained greater than the average dimension (analogous to keeping all factors in a factor analysis with eigenvalues greater than 1), (b) the maximum value of c such that $\sum_t (v_t / V) \cdot 1(v_t / V \geqslant c) \geqslant p$, where $1(\cdot)$ is a binary indicator and p is some selected minimum allowable variance such as $p = 0.50$ (i.e. select c so that at least p percent of the variance is explained), or (c) choosing c such that the number of included dimensions is kept small and easily interpretable (e.g. two or three). We strongly recommend that the ratio of v_t / V be reported and utilized to aid in interpretation, regardless of the selection approach utilized. Let K denote the number of dimensions such that $v_t / V \geqslant c$ $(1 \leqslant K \leqslant T)$, R^* is the reduced rotation matrix with columns r_k^*, and S^* is the reduced PC score matrix with columns s_k^*. Steps 1 and 2 contain the basic analysis steps. The following two steps, which need not necessarily be performed in a specified order, permit inferences from the input: data $\boldsymbol{Y}$, rotation weights R^*, and PC scores S^*.

2.3. Step 3: inference using plots

A number of plots provide information regarding the nature of curve variation. We have found the following three types of plots to be of interest.

(a) Rotation weights vs. measurement index: plot the K curves r_k^* vs. $t = 1, \ldots, T$;
(b) PC scores on different dimensions: plot s_k^* vs. $s_{k'}^*$,
(c) Observed curves for different PC score quantiles: plot y_i corresponding to selected quantiles on $s_1^*, \ldots, s_K^*$.

Plot type (a) is informative in that it provides information about the relative importance of each measurement point t on the kth principal component. Thus, large positive values of r_{kt}^* would indicate that curves with high values of $y_{i,t}$ would tend to have higher scores s_{kt} on the kth PC (and vice versa). For example, imagine a situation in which $r_{11}^* = 1$ and $r_{1t}^* = 0$, $t > 1$. This indicates that the only component of variation among the curves PC 1 is in their initial condition y_{i1}. In Section 3, we will talk more extensively about interpreting these rotation weight curves.

On the surface, plot type (b) seems nonsensical; as by definition, the scores on all pairs of PCs are uncorrelated. However, PCA only implies a lack of *linear* association between their scores and by no means rules out other more complex ones. As we describe below, after we relate the scores on the PCs to "features" of the curves (e.g. initial condition y_{i1} as above) or covariates for a given unit (e.g. income of the household, age of respondent, etc.), identifying pairwise (or more) relationships between PC scores allows us to make statements such as a unit which strongly exhibits one curve feature is more (or less) likely to exhibit a different given feature.

As stated in Section 1, the difficulty in interpreting a plot containing all I unit curves is not in the plot itself, but in the overwhelming number of curves. By using plot type (c), where we select curves to display which lie at different quantiles of each PC dimension (e.g. minimum, median, and maximum), we can display a small number of curves (e.g. three) for each significant dimension. For example, the minimum curve for a given dimension exemplifies a y_i, which has low value on that dimension. Therefore, by span-

ning from lowest to middle to highest values, one might be able to infer, or at least can graphically see, the "cause" (pattern in y_i) which is differentiating units on that dimension. We describe next a more rigorous approach to linking dimension variation to curve features.

2.4. Step 4: linking variation in dimensions to covariates using regression

Our interest lies in explaining variation in curves due to the curve features as well as other covariates (not included in the PCA analysis). For each curve y_i, compute curve features of interest and obtain unit covariates $x_{i1}, x_{i2}, \ldots, x_{ip}$. For example, if each curve is weekly purchase quantity (say in units) for a given SKU, then features of interest might include week of initial purchase, quantity of initial purchase, total purchases, etc. In addition, auxiliary covariates collected on each household, such as income, number of household members, and region of the country, may also drive curve variation and be of interest. Then, one analysis of interest is to correlate each vector of PC scores, s_k^*, with the features and covariates x_{ip}, and the features and covariates with each other. This can be performed using a standard simple univariate regression routine or via a correlation calculation. An analysis of a feature PC score correlation could possibly inform, above and beyond the plot of y_i for different quantiles of s_k^*, the features of the curves that are driving variation in the PC dimensions. The correlation among the feature scores is of interest in two respects: (1) it is informative to see interdependencies among features of curves (e.g. units with large total volume purchased may also have a high number of initial purchases), and (2) it may suggest more complex relationships about curve variation. In this respect, it may be informative to run a multivariate regression of each vector of dimension scores s_k^* on the curve features and covariates; key determinants of variation are those features and covariates found to be significant at a prespecified level.

3. Simulation

To demonstrate our approach, we simulated curves exhibiting three common types of variation observed in marketing: unit heterogeneity, time-varying parameters, and seasonal spikes. Specifically, and without loss of generality, each simulation was done using $I = 1000$ units (curves) and $T = 52$ measurements (e.g. weeks). Other simulation structures of interest are briefly described in Section 5.

3.1. Simulation 1: unit heterogeneity

A total of 500 curves were generated by simulating 52 observations from a Gaussian distribution with mean $\mu = 4$ and variance $\sigma^2 = 1$. A second set of 500 curves were obtained from a Gaussian with mean $\mu = 3$ and $\sigma^2 = 1$. The means were chosen so that there would be significant overlap among the two groups to both see how the approach would perform, and also that a simple graphical display would not lead to trivial inferences (as would be indicated if the separation was large). This situation can approximate a market in which there are two segments: heavy and light users, with stationary purchase behavior. A plot of the 1000 curves is given in Fig. 1, panel (a). Clearly, the degree of overlap and physical volume of ink is so large as to make this plot of little use.

A PCA was run on the data; not surprisingly, only one significant PC was found accounting for 22.0% of the variation in the curves (the second PC explained 2.0% of the variation). A plot of the PC rotation r_1^* vs. measurement index $t = 1, \ldots, 52$, PC 1 scores s_1^* vs. PC 2 scores s_2^*, and y_i corresponding to the minimum, median, and maximum scores on PC 1 are given in Fig. 1, panels (b), (c), and (d), respectively.

In panel (b), we see that the PC 1 rotation is constant across measurements (mean $= -0.14$), i.e. each observation is equally important (weighted) in determining variation in the curves. This is as expected, as the heavy and light users were generated according to a constant mean stationary process. As a general result, straight horizontal line PC rotations (at constant nonzero values) imply that the PC is related to total volume, that is

$$s_i = \sum_{t=1}^{T} \text{constant} \cdot y_{it} = \text{constant} \cdot V_i, \qquad (1)$$

where V_i is the total "volume" for unit i.

In panel (c), the plot of PC 1 and PC 2 scores clearly demonstrates the separation of the two seg-

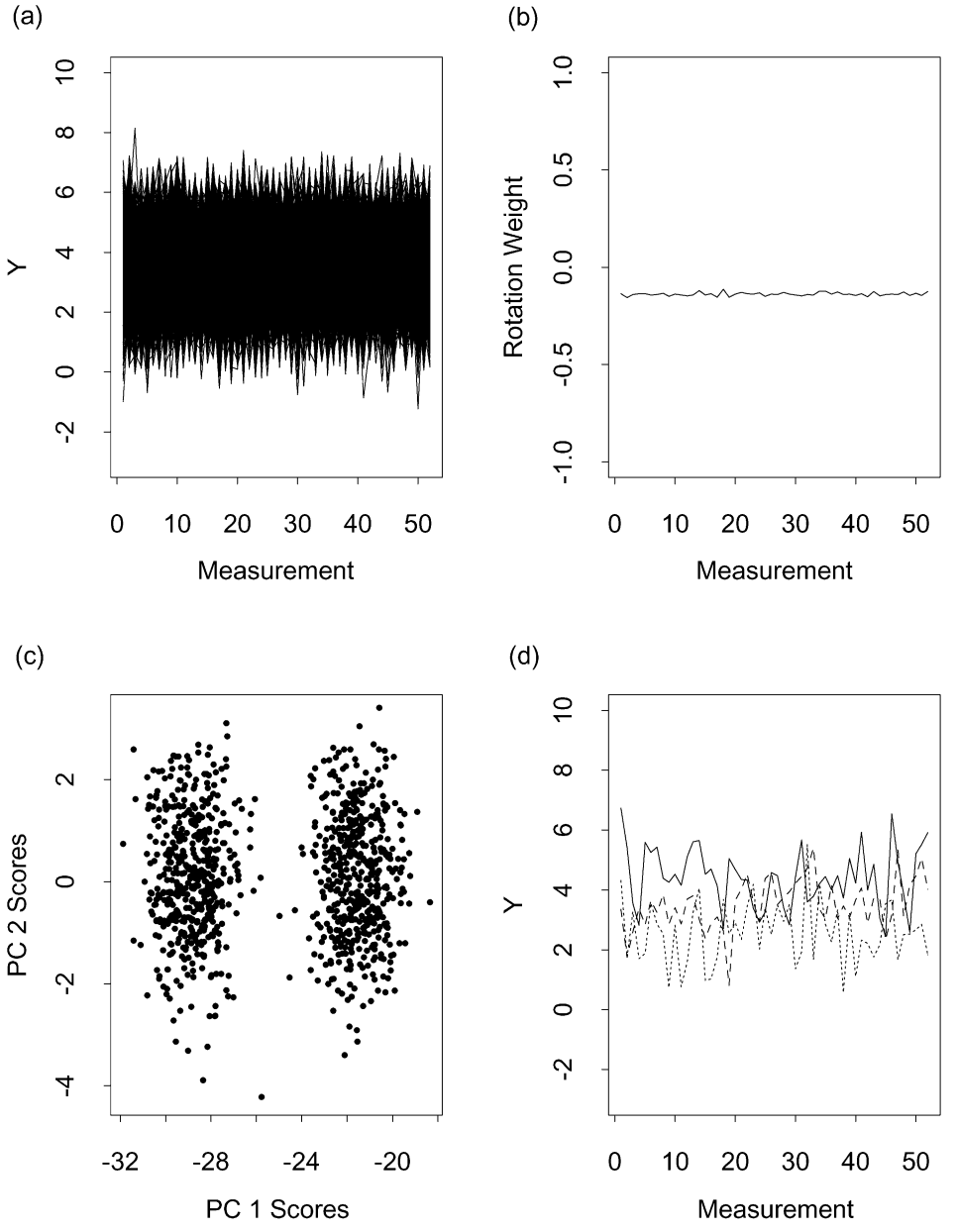

Fig. 1. Results for Simulation 1: unit heterogeneity. Panel (a) contains a plot of the 1000 curves vs. *t*. Panel (b) contains a plot of the PC 1 rotation vector vs. *t*. Panel (c) contains a plot of PC 1 vs. PC 2 scores. Panel (d) contains a plot of the curves for the minimum, median, and maximum units on PC 1. The minimum curve is a solid line; the median curve contains long dashes and short dashes in the maximum curve.

ments (one segment's PC 1 mean is at 21.8 and the other at 29.1) and additionally, the marginal independence of PC 1 and PC 2. A histogram of the PC 1 scores (or of the initial conditions y_1), not shown, would also demonstrate this bimodality and indicate that 500 curves are in each segment. Furthermore, from panels (b) and (c), we can also infer that one segment has a mean around 3 and the other segment

around 4 (which we know is true), by utilizing Eq. (1), setting the constant = − 0.14 (the PC 1 rotation mean from panel (b)), and solving for the two segment volumes V_i where we set the mean $s_i = 21.8$ for one segment and 29.1 for the other (note that the difference in means equal to $1 = -0.14 \times (29.1 - 21.8)$ is also implied).

Fig. 1, panel (d) of the minimum, median, and maximum PC 1 curves provides a number of interesting insights. First is that the lowest, median, and highest curves have comparable variances and shape. In particular, stable behavior over the entire T measurements is indicated, with most of the variation due to an intercept shift. No spikes or other aberrancies are apparent.

We also report on a power study that was performed using simulation by varying the signal-to-noise ratios (mean difference to variance) in the two groups (as in Simulation 1, i.e. a mean shift). This was done by running 1000 simulations per condition with two-sample t-statistics (labeled true t in the table below) equal to 0 (no effect), 1 (modest effect), 2 (strong effect), and 3 (almost separable) between the groups. After simulating the data, we designated that the PCA approach "found the pattern" if: (1) the first PC vector had $v_1 > kV$, i.e. the first PC had a large fraction of the total variance explained, (2) it corresponded to the constant term (established by computing the variance of rotation vector for PC 1), and (3) no other significant (above average) PC was found. We chose $v_1 > kV$ with $k = 0.025$, 0.05, and 0.10, as it represents a PC with roughly 2.5, 5, and 10 times the average PC variance. Power calculations for different ratios are also feasible; however, this spanned a reasonable range.

The results of the power study are summarized in the Table 1 where we report the percentage of the 1000 simulations that the PCA approach deemed a significant PC 1 constant term for various values of k as well as other simulation statistics for the first PC variance v_1.

Table 1

True t	Average v_1	v_1/V	$k = 0.025$	$k = 0.050$	$k = 0.10$
0	1.48	0.028	1	0	0
1	3.79	0.072	1	1	0
2	7.31	0.130	1	1	1
3	10.91	0.182	1	1	1

We note a number of expected patterns. First, as the true t increases, the average value of v_1 and ratio of v_1/V increase. We also note that setting of k too low can be of concern. For example, when there is no effect (true $t = 0$), all 1000 of the simulations had *largest* PC variance v_1 greater than 2.5 times the average (this is not always true and depends upon the variances of the two groups). This is due to the fact that the maximum of a set of "uniformly" distributed PC values had a maximum greater, in this case, than 2.5 times the average. Finally, the v_1/V column provides some guidance as to how large the percentage of variance explained by the first PC will be, which can be used as a baseline to set k in practice. In summary, as we suggest in Section 5, this type of simulation, in accordance with using this approach in practice, is recommended.

3.2. Simulation 2: time-varying parameters

A significant amount of research in the recent marketing literature has been about models which allow their parameter's values to dynamically change over time (e.g. Fader & Lattin, 1993). We simulated a dynamically varying structure in the following manner. A total of 250 of Simulation 1's units with mean 4 and variance 1 were redrawn from a Gaussian distribution with mean $3.5 + t/52$ and variance 1, i.e. their means vary linearly with time, but average the same as the $N(4, 1)$ group. In addition, 250 of the units with mean 3 and variance 1 were redrawn with mean $2.5 + t/52$ and variance 1 (identical mean to $N(3, 1)$). Notice now that there are four simulated groups in a 2×2 design (high vs. low mean and linearly increasing vs. constant mean over measurements): $N(3, 1)$; $N(2.5 + t/52, 1)$; $N(4, 1)$; $N(3.5 + t/52, 1)$; where the first two groups have the same overall mean, but different shapes (similarly for the last two groups). Under this simulation structure, we expect two significant PCs, one for overall volume and one for shape. A plot of the 1000 curves is given in Fig. 2, panel (a).

A PCA analysis of these 1000 curves did indeed indicate two PCs which explain a large proportion of the variance (72% of the variance explained on the first PC and 22% on the second); their rotation vectors are plotted in Fig. 2, panel (b). We note that the first PC is indeed a straight horizontal line which differ-

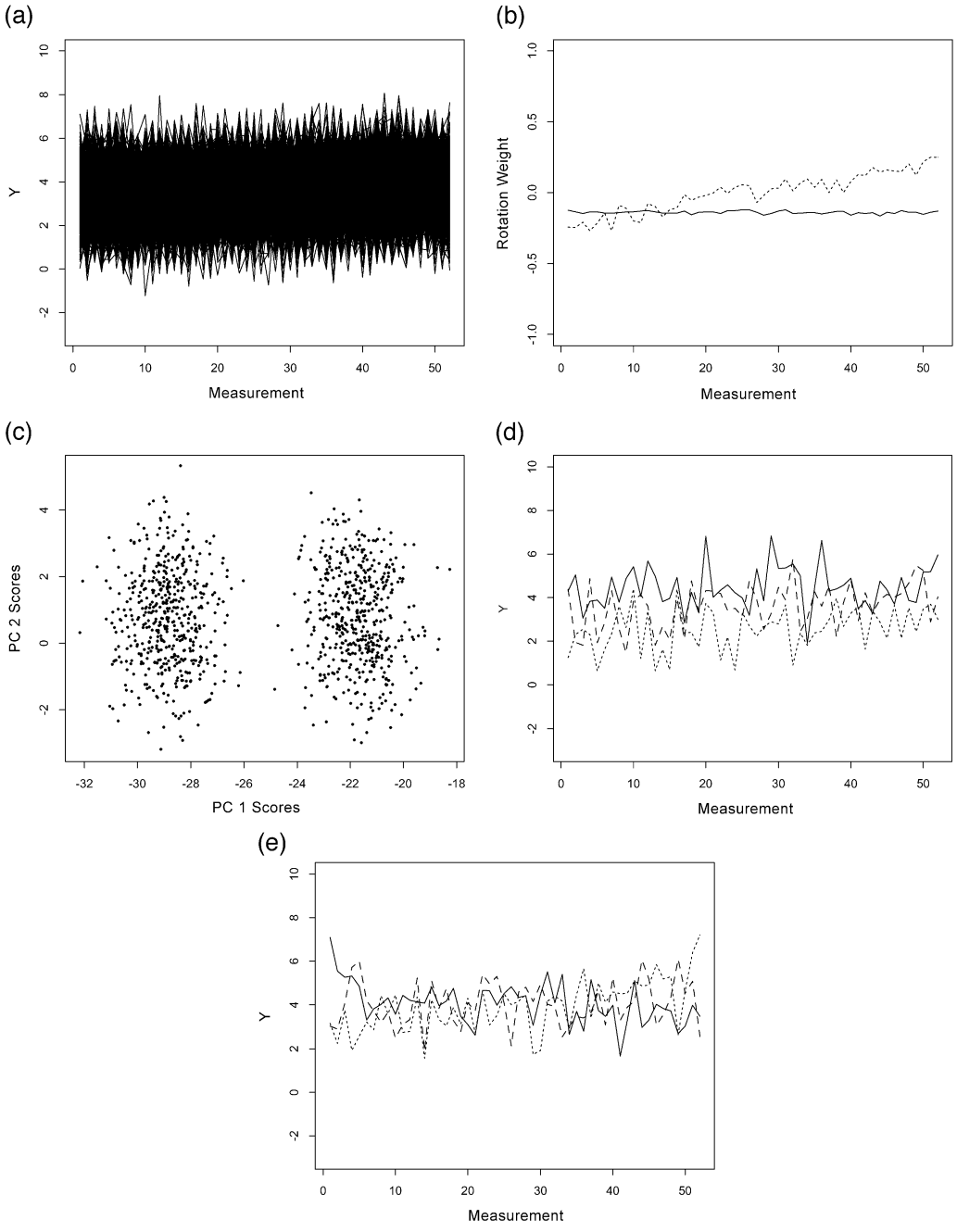

Fig. 2. Results for Simulation 2: time-varying parameter. Panel (a) contains a plot of the 1000 curves vs. *t*. Panel (b) contains a plot of the PC 1 and PC 2 rotation vector vs. *t*. The solid line is the PC 1 vector, the dashed line is the PC 2 vector. Panel (c) contains a plot of PC 1 vs. PC 2 scores. Panel (d) contains a plot of the curves for the minimum, median, and maximum units on PC 1 and panel (e) for PC 2. The minimum curve is a solid line; the median curve contains long dashes and short dashes in the maximum curve.

entiates the two groups with mean equal to 3 from those two where the mean equals 4. The second principal component is linearly increasing, showing the separation of the groups with the same means, but different shapes. Furthermore, the PCA indicates that the optimal differentiation is (and accurately so) linearly increasing with t. It is also of importance to note that the ratio of 72% variance on PC 1 to 22% on PC 2 is driven by the mean difference between groups 1, 2 and 3, 4 compared to the slope of the linear trend (1/52). A greater slope would have yielded a higher effect due to PC 2 (and in fact could become PC 1 for an even greater slope).

In Fig. 2, panel (c), we plot the scores on PC 1 vs. PC 2. The separation of the two groups with mean 3 and two groups with mean 4 is evident from the marginal distribution on PC 1. The separation of the groups on PC 2 is less evident (again, a steeper slope would have yielded a more dramatic effect), graphically due to the overlap between the groups; however, under inspection of the scores themselves, this is readily apparent.

In Fig. 2, panels (d) and (e), we show the minimum, median, and maximum curves on PC 1 and PC 2, respectively. Notice in panel (d) that the minimum curve was from the $N(4, 1)$ group; this was purely by simulation error as the $N(3.5 + t/52, 1)$ group has the same mean on PC 1 and, in fact, the second smallest PC 2 curve corresponded to one from this group. A similar result holds for the maximum PC 1 score and the two groups with mean equal to 3.

A correlation analysis indicated that PC 1 was perfectly negatively related to total volume purchased ($r = -0.9999$, $p < 0.0001$); this is not surprising given the horizontal line indicated in Fig. 2, panel (b). In addition, PC 2 is perfectly related to the linear increasing trend in y_i vs. t ($r = 0.9999$, $p < 0.0001$). One thing to note is that from the other plots, the feature "linearly increasing" becomes evident and hence makes it easy to explain PC 2 in this case. However, in some cases in which the plots do not inform the user about which curve features have value, it may not be trivial to "explain" the variation in the PC scores; yet, a regression on all available features/covariates is a good starting place.

As a power study in this simulation case, and one for the next simulation provided "identical" results to that given for Simulation 1 (as expected given the linear nature of all three simulation effects), we do not report further on their findings.

3.3. Simulation 3: seasonal spike

Periods of extremely high purchasing activity (spikes) in time series, reflecting holiday period buying behavior, are commonly observed. We assessed the ability of the PCA approach to uncover spikes by simulating 500 curves with $y_{it} \sim N(3, 1)$ and 500 curves with $y_{it} \sim N(3, 1)$ for all weeks excluding $t = 51$ (Christmas week), and $y_{it} \sim N(6, 1)$ for $t = 51$ that is double the normal average buying behavior during the holiday period. Plots of the 1000 curves, first PC scores vs. week, PC 1 scores vs. $y_{i,51}$ (week 51 measurements) and minimum, median, and maximum PC 1 curves are given in Fig. 3, panels (a), (b), (c), and (d), respectively.

From panel (a), the spike at $t = 51$ is observable, yet somewhat hidden by the large set of curves. In addition, the significance of the spike is mostly masked by the plot. Of course, the set of means by week would identify the spike at week 51 as well. In panel (b), we clearly observe the PC 1 spike at the $t = 51$ measurement (PC 1 accounts for 6.5% of the curve variations) and suggests panel (c) in which we plot the PC 1 scores vs. $y_{i,51}$. From panel (b), we also note that the depth of the spike is at -1, indicating that the variation on PC 1 is essentially entirely due to week 51. We also note that a larger fraction of the curve variation (greater than 6.5%) would occur if the signal (mean of $y_{i,51}$ for segment 2) was made greater or the noise (variance) of the groups was reduced. The plot in panel (c) indicates the almost perfect correlation ($r = -0.995$, $p < 0.0001$) between PC 1 score and week 51 measurements. The minimum, median, and maximum PC 1 curves given in panel (d) further demonstrate PC 1's link to the height of the spike (the minimum, median, and maximum curves have spike height at 9.73, 4.36, and -0.10, respectively).

4. Real data example

We applied our PCA approach to a real data set containing the cumulative penetration rates obtained from market tests conducted by Information Resour-

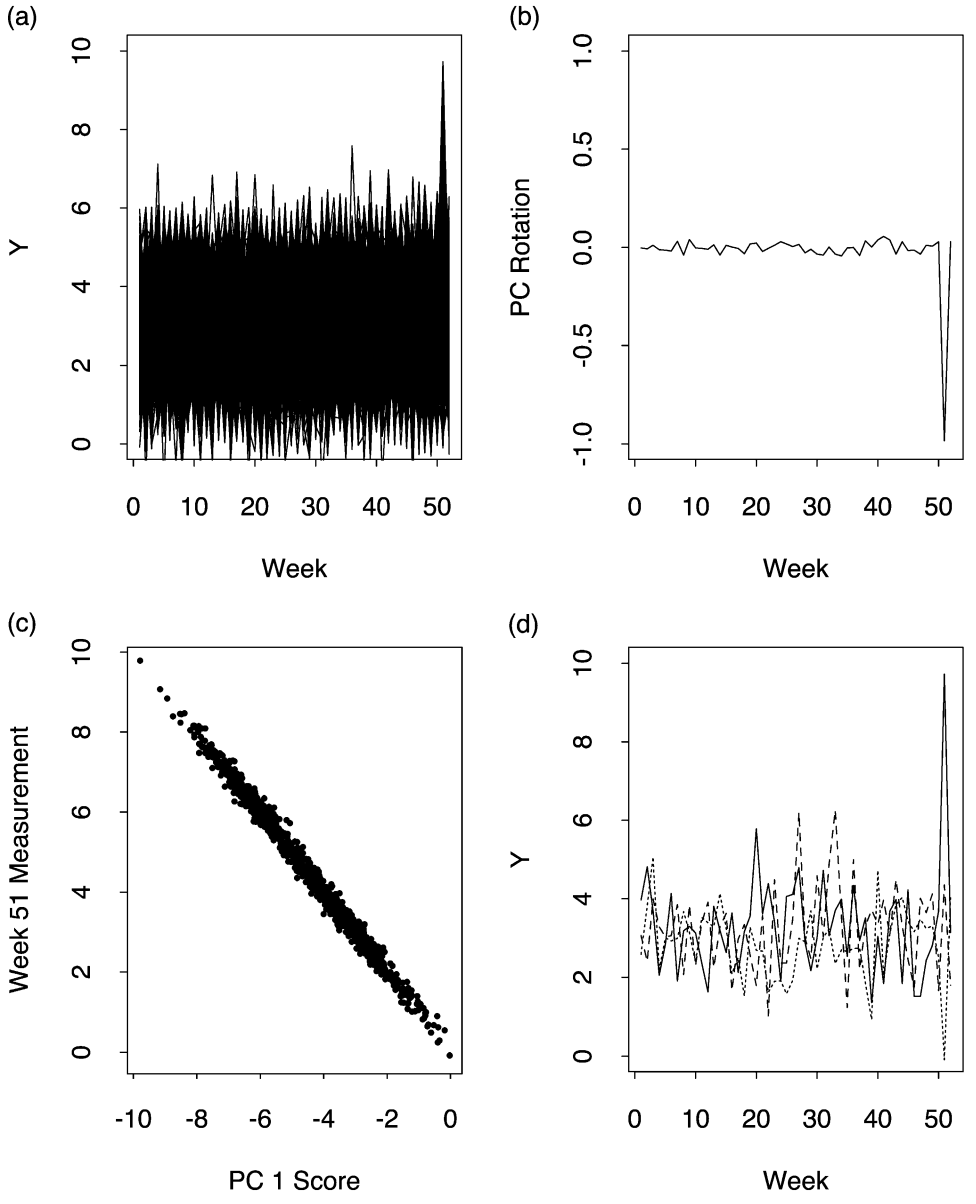

Fig. 3. Results for Simulation 3. Panel (a) contains a plot of the 1000 curves vs. *t*. Panel (b) contains a plot of the PC 1 rotation vector vs. *t*. Panel (c) contains a plot of PC 1 scores vs. week 51 measurement. Panel (d) contains the minimum, median, and maximum PC 1 curves. The minimum curve is a solid line; the median curve contains long dashes and short dashes in the maximum curve.

ces' Behavior Scan consumer panel for 19 products over 52 weeks ranging from 1989 to 1996. This is the data set analyzed in Hardie, Fader, and Wisniewski (1998), hereafter HFW, in which they compare the fits of eight different diffusion models using three different estimation techniques. Diffusion data sets are a natural candidate for our approach due to their longitudinal structure and the intrinsic interest in recent research to understand variation in shapes of diffusion curves across products, countries, and time (Van den Bulte, 2000). Thus, in general, by implementing our PCA approach and regressing the PC scores on curve

covariates such as country of origin, date of introduction, or type of product (see Step 4 in Section 2), preliminary understanding into diffusion rate variation can be obtained.

4.1. A product diffusion data set

A PCA was run on the HFW data set for which the 19 cumulative product diffusion curves are shown in Fig. 4, panel (a). Here, with only 19 curves, obtaining "a few basic curve shapes" is visually feasible; yet, our PCA approach can still provide further insight. Some interesting features of the data to note are the one outlying curve which had fast take-off, yet slow-growth thereafter, and the significant variation among the curves in take-off rate and subsequent growth. An analysis was performed with the outlying curve removed; yet, no significantly different substantive findings were obtained and, hence, we report the analysis based on the full set of 19 products.

Fig. 4, panel (b) contains a plot of the first two PC rotation vectors accounting for 88% and 10% of the curve variation, respectively. The first PC rotation, which is approximately a constant horizontal line, indicates (as mentioned earlier) that PC 1 is essentially total volume, or in this case, cumulative penetration at week 52. This is corroborated (via a Step 4 analysis) by a highly significant correlation of $r = -0.988$ ($p < 0.0001$) between PC 1 scores s_{i1}^{*} and week 52 cumulative penetration $y_{i,52}$. Thus, as expected, the greatest variation in the curves can be summarized by their variation in total success over their target periods. The second PC rotation, which starts out negative and goes to increasingly positive values, is therefore related to the "shape" of the curve, being maximized by a product which would start out slowly and take a late rise. These two PCs tell an interesting story. The first PC would be high for a product which takes off fast (large innovator group) and the second for a product which takes off late (large imitator segment).

A plot of the PC 1 vs. PC 2 scores is given in Fig. 4, panel (c). No obvious pattern is exhibited among the two dimensions; however, it is interesting to note that the product which is lowest on PC 1 is low on PC 2; yet, the product which is highest on PC 2 is relatively low on PC 1. In addition, we can see that there is an outlier with respect to PC 1 (total volume). In this case, this is evident from panel (a), the connected symbol plot of the raw data. However, in many cases, visually observing outlying curves is impossible and the PC scores facilitate their identification. Gnanadesikan and Kettenring (1972) discuss, in general, the use of various multivariate techniques to identify outliers. It is important to note that the PC scores not only aid in the identification of outliers, but also in stating their direction (i.e. why are they outliers). Here, we see that the curve is outlying with respect to PC 1 which we know is total volume. Thus, the combination of outlier detection and "explaining" their reason provides a powerful managerial tool obtained via the PCA approach.

The relationship between PC 1 and PC 2 is further investigated in Fig. 4, panels (d) and (e). In panel (d), we plot the minimum, median, and maximum products on PC 1, clearly demonstrating the relationship between PC 1 and week 52 cumulative penetration. The minimum product, the outlier, which has an extremely fast take-off, has the largest cumulative penetration making it high on PC 1; yet, its fast take-off makes it a "curve shape" low on PC 2. Panel (e) contains the minimum, median, and maximum products on PC 2. It is of interest to note that the median curve is not bounded by the minimum and maximum ones, further corroborating that PC 2 is not measuring cumulative penetration as is PC 1. In fact, $r = 0.145$ for PC 2 and week 52 penetration which is not significant, but either way is in the opposite direction of the correlation from PC 1. A more detailed analysis of the maximum curve from PC 2 in panel (e) indicates that this product had only the 8th, 13th, and 14th largest penetration as of weeks 1, 5, and 10, respectively (a moderate take-off), but by week 52, had achieved the second highest cumulative penetration rate. In contrast, the minimum curve had the 12th, 10th, and 9th highest penetration rate as of weeks 1, 5, and 10, respectively, and ended up the 7th highest; so this product actually lost ground comparatively.

In summary, for this diffusion data set, the PCA approach has been able to uncover two major sources of variation: overall penetration and "late bloomers," which were not immediately visually evident and correspond to established theories of innovators and laggards.

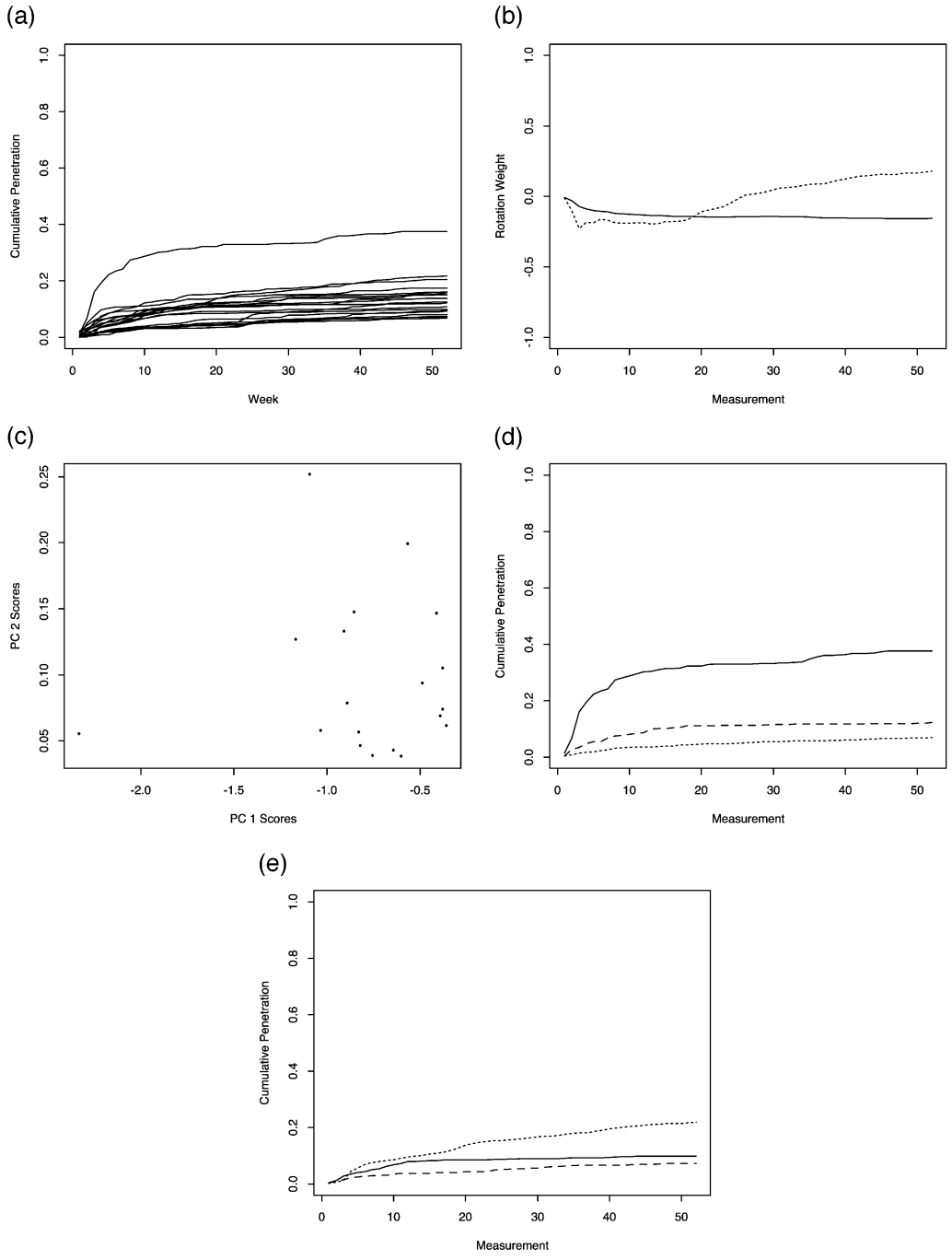

Fig. 4. Results for Example 1. Panel (a) contains a plot of the 19 cumulative penetration curves vs. *t*. Panel (b) contains a plot of the PC 1 and PC 2 rotation vector vs. *t*. The solid line is the PC 1 vector, the dashed line is the PC 2 vector. Panel (c) contains a plot of PC 1 vs. PC 2 scores. Panel (d) contains a plot of the curves for the minimum, median, and maximum units on PC 1 and panel (e) for PC 2. The minimum curve is a solid line; the median curve contains long dashes and short dashes in the maximum curve.

5. Summary and conclusions

In this research, we have presented an approach to understanding basic sources of variation in repeated measures data sets which does not require running a formal model. Moreover, the approach can be applied very quickly, runs in standard software, and utilizes techniques that analysts and managers commonly employ. We have provided simulations, which indicate that the approach can identify latent segments, time-varying heterogeneity, and seasonal spikes of a specific form and, when relating the heterogeneity found to known features and covariates, can provide interesting insights. Furthermore, by looking at each curve summarized by a point on an underlying dimension, identifying outliers is made possible; moreover, the user can say that the point is outlying with respect to that PC dimension. This can be of great managerial interest as now not only can a unit's outlyingness be identified, but also the "direction" stated, e.g. the unit is outlying because its total volume (PC 1) is low.

However, by no means do we imply that this approach is foolproof. First, interpretation of the output, plotting the appropriate curves, computing the correct curve features, etc., still require careful thought; possibly even more so than for formal models in which inferences can be derived directly from parameter estimates. For this reason, we strongly recommend that in conjunction with this approach, the analyst considers a simulation (possibly by bootstrapping) to evaluate the efficacy of the approach.

Secondly, the PCA approach is likely to have low power in detecting certain structural types of variation. For example, curves in which higher order moments of their shape distinguish them can easily be missed. In fact, we ran a simulation with 500 curves obtained from a $N(3, 1)$ and 500 curves obtained from a Gamma(9, 3), which have the same mean and variance, but different structural variation. The PCA approach was unable to find this structure, albeit this is a very "difficult" test to pass.

Thirdly, there are practical issues in applying the method to data. For example, for the product penetration data, "Do we run the analysis on incremental or cumulative penetration rates?" Running the analysis on cumulative data was done, as the incremental weekly figures were noisier and led to PCs with less straightforward interpretation. However, this certainly can effect what is found.

In summary, using standard software to run exploratory analyses beyond basic plots, crosstabs, means, and variances can provide many insights. Our hope is that for data sets of this type, people will adopt our approach as part of their everyday toolkit. Even for the staunch modelers among us, it can provide direction for determining complex models and possibly even inform about their necessity before investing time in them.

Acknowledgements

The author thanks Peter S. Fader for significant contributions on earlier versions of this work and the editor and three anonymous reviewers for their suggestions.

References

Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.

Fader, P. S., & Lattin, J. M. (1993, Summer). Accounting for heterogeneity and nonstationarity in a cross-sectional model of consumer purchase behavior. *Marketing Science*, *12*(3), 304–317.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*, 81–124.

Hardie, B. G. S., Fader, P. S., & Wisniewski, M. (1998). An empirical comparison of new product trial forecasting models. *Journal of Forecasting*, *17*, 209–229.

Jain, D., Bass, F. M., & Chen, Y. (1990, Feb). Estimation of latent class models with heterogeneous choice probabilities: an application to market structuring. *Journal of Marketing Research*, *27*(1), 94–101.

Jones, M. C., & Rice, J. A. (1992). Displaying the important features of large collections of similar curves. *American Statistician*, *46*(2), 140–145.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Ramsey, J. O., & Silverman, B. W. (1997). *Functional data analysis*. New York: Springer-Verlag.

Rossi, P. E., & Allenby, G. M. (1993, May). A Bayesian approach to estimating household parameters. *Journal of Marketing Research*, *30*(2), 171–182.

Russell, G. J., & Kamakura, W. A. (1994, May). Understanding brand competition using micro and macro scanner data. *Journal of Marketing Research*, *31*(2), 289–303.

Tucker, L. (1958, March). Determination of the parameters of a

functional relation by factor analysis. *Psychometrika*, *23*, 19–23.

Tucker, L. (1968). Comments on confounding sources of variation in factor analytic techniques. *Psychological Bulletin*, *70*(5), 345–354.

Tucker, L., & Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika*, *28*, 333–367.

Van den Bulte, C. (2000, Fall). New product diffusion acceleration: measurement and analysis. *Marketing Science*, *19*(4), 366–380.