

Modeling Dynamics in Online Auctions: A Modern Statistical Approach

Galit Shmueli & Wolfgang Jank
Dept of Decision & Information Technologies
Robert H. Smith School of Business
University of Maryland, College Park

Abstract

In this work we propose a modern statistical approach to the analysis and modeling of dynamics in online auctions. Online auction data usually arrive in the form of a set of bids recorded over the duration of an auction. We propose the use of a modern statistical approach called functional data analysis that preserves the entire temporal dimension in place of currently used methods that aggregate over time thereby losing important information. This enables us to investigate not only the price evolution that takes place during an auction, but also the *dynamics* of the price evolution. We show how functional data analysis can be combined with cluster analysis and regression-type models for data exploration and summarization, and for testing hypotheses about relationships between price formation and other relevant factors.

Keywords: Functional data analysis, price evolution, eBay, smoothing, clustering, regression

This research was partially funded by the NSF grant DMI-0205489

1 Introduction and Motivation

There is a growing body of empirical research in the fields of economics and information systems that is concerned with the study of online auctions. A variety of questions are investigated, among them: What factors affect the final price of an auction? Why does last-minute-bidding occur and why do people bid early in the auction? Is the phenomenon of Winner's Curse present in online auctions? Studies that employ statistical tools for answering such questions typically use a single measure, such as the final price, or total number of bids as the dependent variable of interest. These variables are static in the sense that they only give a snapshot taken at one particular time point, usually at the end of the auction. However, they ignore the entire duration of the auction, and in particular the *dynamics* of the price formation or bidding process as it evolves throughout the auction. In this paper our aim is to study the entire price formation process and its dynamics in online auctions. By "price dynamics" we mean the progress of the price throughout the auction, its speed, changes in speed, etc. An analogy is cars in a car race. Instead of focusing exclusively on the winner's score (or the winning price in an auction) as the dependent variable, we look at the route, speed, acceleration, and other dynamic characteristics specific to the choices the driver made. Clearly, there is a relationship between these dynamics and the winner's score. However, the race dynamics are interesting in themselves and can be useful for understanding other important phenomena. For instance, better acceleration performance of our driver can be an indication of a superior car. Thus, the dependent variable from our viewpoint is not a static point variable, but rather the price curve throughout the auction. We investigate how price increases throughout an auction: How fast does it increase at different stages of the auction? How fast does it move towards the final price? Which dynamics are common and which are different across various auctions? We also study factors that influence the price formation, in particular the minimum price (which is analogous to a reserve price in auction theory, [3]). Finally, relationships between the auction dynamics and other variables of interest such as final price are also investigated.

From a data structure point of view, static variables call for aggregated data whereas dynamic data use the entire un-aggregated sequence of bids. For this reason previous studies tended to aggregate available information over the auction duration, over auctions, or over both. In the first

case, an entire auction is reduced to a single time-point, usually the end. Examples are modeling the closing price (a single point for each auction) as a function of various factors such as the number of bids, the auction duration, and the seller rating [17]. Aggregation over auctions is the case where the data from a set of auctions are treated as one set of datapoints. For example, [28] look at the empirical cumulative distribution function of bid times that are taken from multiple auctions but they combine them into a single sample. Finally, aggregation over time *and* auctions occurs when the object of interest is, for example, the final price, and all final prices are aggregated, so that they are represented by their mean. This type of double-aggregation is wide spread, and typically appears as a table of means. Although aggregation is convenient for summarizing data, it carries the risk of losing important information to the degree of arriving at wrong conclusions (this is also known as “Simpson's paradox”). Once data are aggregated, the information loss is usually unrecoverable. Clearly, it would be ideal to explore and model the information contained in the entire duration of each auction, and to aggregate as late as possible in the modeling process, and only when deemed necessary. One of the main reasons that such analyses are absent from the online auction literature is most likely the inadequacy of popular methods such as regression models and time series analysis for fitting such data. Online auction data have a special form that is not traditional in the statistics literature. On the one hand, the bidding sequence in a single auction forms a time series with special features (such as unequally spaced observations over a finite interval). On the other hand, we have not a single series but multiple time series, which are not “aligned” or of the same length. Such a data structure is not common and thus commands special statistical methods.

It appears that there is a gap between online auction researchers, who are typically from the fields of economics and information systems, and researchers in the area of statistics. In part this is due to the new data collection mechanisms that have not swept the statistics academic community: In order to obtain online auction data (and other e-commerce type data) software agents are commonly used. The software agents, also known as spiders, are typically written by the research team and the code is specific for the application at hand. [13] discuss the implications of these modern data collection mechanisms and how their availability enables empirical research that has not been possible earlier¹. Statisticians usually do not have the knowledge or expertise to write such agents, and are therefore left outside of this rich world of data and the

research opportunities that they create. This gap between disciplines certainly calls for collaboration between data collection gurus and data analysis masters!

Our main goal is to show how statistical thinking and modern statistical methodology can be useful for exploring, gaining insight into, and testing hypotheses with such data, in the sense of capturing processes and their dynamics. In Section 2 we describe the special features and structure of online auction data, which makes traditional statistical analysis methods such as regression and time-series models less adequate for exploring their dynamics. We also discuss the data structure from a statistical perspective and define the sample-population relationship of interest. We then introduce the method of Functional Data Analysis (FDA): we explain how FDA differs from classical statistical methods, and why we think it is suitable for exploring and analyzing online auction data. Table 1 summarizes the main differences between a “static” approach and our “dynamic” approach in terms of data structure, research questions, and statistical models.

	Static (point) response	Dynamic curve response
Examples	final price, number of bids, number of bidders, average bidder rating	sequence of bids over time, cumulative number of bid, number of distinct bidders over time, bidder rating over time
Data structure	single measurement per auction	time-series for each auction
Research question format: How are factors related to...	increase/decrease in response	shape, speed, acceleration of response curve
Typical summary statistics	averages, percentiles, counts	average and percentile curves
Typical plots	histograms, scatter plots, bar charts	profile plots, phase-plane plots
Statistical models	regression-type models; cluster analysis	functional regression-type models; functional cluster analysis

Table 1: Comparison of a traditional “static” dependent variable with a dynamic one

In Section 3 we illustrate how the method works using real data from eBay.com. We perform exploratory analysis and confirmatory analysis of the price formation process and tie our results

with auction theory and the current literature on online auctions. We describe the different steps involved in FDA, and show what type of information and knowledge are gained at each step. Thus, our goal is to integrate research question, statistical methodology, and data within this expository work. In section 4 we discuss additional aspects of auction dynamics and suggest future directions for applying modern statistical methods to auction data and more generally to e-commerce data.

2. Data Structure and Statistical Setting

2.1 Features of Bid histories

Bid data from online auctions are very different from data used in traditional statistical analyses in several respects. In the following, we use the term “bid history” to denote the sequence of bids and time stamps for a single auction. In that sense, a bid history is a time series. In closed-end auctions such as eBay, the auction length is predetermined by the seller, and thus the time series takes place over a preset, finite interval. In open-ended auctions the duration could be, at least in theory, infinite. We focus here on closed-end auctions only.

Unlike traditional time-series analysis, where the problem at hand is to fit a model or forecast a single time series, the goal in analyzing auction data is to gain a better understanding of other (current or future) auctions that are not part of the dataset. In contrast to traditional time series analysis which is concerned with a single time series at a time, we have multiple (typically several hundred) such series. Furthermore, in a single time-series analysis the unobserved population of interest is usually the future (i.e., times $t+1$, $t+2, \dots$), whereas in the auction setting it is the unobserved population of all similar auctions that are not in the dataset (and not the “future” since the auction is closed!). Thus the goals, the data structure, and the statistical setting are different from ordinary time series analysis.

The three features that set bid histories apart from traditional time series are the unequally spaced observations within each auction, the different number, and the different time stamps across different auctions. Even if we “standardize” all the auctions to start at time 0 and end at time T , we have multiple time series with varying numbers of observations, which are unequally spaced and differ across auctions. This means that even from a data entry point of view we cannot store the data easily in a traditional matrix form. This data structure is common in various online auction formats, including reverse auctions.

2.2 Statistical Framework

In light of the objectives of our study, we define what we mean by a sample, by a population, and the relationship between the two. Treating each bid history as a time-series, we view our sample

of auctions as a sample from the entire population of all relevant auctions. Thus, our observation of interest is not a scalar, but rather an entire function. We then use the sample information to investigate the dynamics of the entire population, in the sense of the price evolution throughout the auction, its speed, and more. In addition to exploring patterns of price dynamics, we also examine heterogeneity across auctions and aim to find factors that drive such heterogeneity. In particular, we strive to compare the price dynamics in auctions either for the same item, or even across auctions for different types of items (electronics, collectibles, etc.). On the one hand, it is possible that there are different patterns of price dynamics even among auctions for the same item. On the other hand, although prices are expected to be different across different items (for example, between a DVD and a collectible coin), the dynamics of the price formation can be similar. Thus, the population of all auctions need not necessarily be partitioned into sub-populations according to the item sold.

We approach the data analysis task from a statistical angle: we start by simple exploration before moving to more formal confirmatory analysis. The goal of the exploratory phase is simply to display the data for the purpose of familiarization with its structure, features, and complexity - a very important step which is often overlooked. At this point, we do not yet test hypotheses. We use graphical displays, summarization, and data reduction techniques. Since the data have special features, ordinary displays such as histograms, scatter plots, and bar charts lead to a loss of the temporal information. We thus strive to limit the loss of information as much as possible.

In many papers that analyze online auction data, aggregation of bids across time appears to be a commonplace practice (e.g., [2]). This means that instead of viewing an auction over its entire duration, it is reduced to merely a single time point. The dynamics of the auction, as it turns out though, are important for capturing, understanding, and evaluating phenomena that are found in the aggregated data. One example is the phenomenon of "last minute bidding", which has been examined through looking at the number of bids during the last minute or so of the auction and comparing it to the overall number of bids in the auction ([2], [32]). An alternative, which preserves the temporal information, is to look at auctions as time-series. [30] developed a three-stage model that describes the bid arrival process in closed-end auctions such as eBay. They treat the incoming bids as points in a continuous process, thereby maintaining the temporal

information. Applying the model to real data, they show that the “last minute bidding” actually occurs over the last couple of minutes, and that the bid arrivals are actually more moderate than expected in light of the rapidly increasing bid arrivals until the last 2 minutes of the auction.

2.3 A Suitable Statistical Method: Functional Data Analysis

Since our object of interest is the sequence of bids (representing the price curve) throughout the entire auction, we choose to use the method of functional data analysis (FDA) for exploring and analyzing online auction data. FDA is a cutting edge statistical method which was made popular by the monograph of [24] and earlier work by the two authors and others. In FDA, the object of interest is a set of curves, shapes, objects, or, more generally, a set of functional observations, rather than a set of data points, as it is the case in classical statistics. It therefore enables the representation and analysis of observations where the underlying multiple measurements per observation are realizations of a single object. Unlike multivariate methods that treat the multiple measurements as separate entities, in FDA the assumption is that the multiple measurements form a continuous object, like a curve. FDA has gained momentum in various fields of application such as the agricultural sciences [18], the behavioral sciences [27] as well as medical research [20]. The method has been used to analyze the dynamics of seasonally-varying production indices [23], to predict El Nino [6], and to study the dynamics of interest rate curves [19]. An excellent collection of case studies involving functional data analysis can be found in [25].

Unlike longitudinal methods that are suitable when the number of observations per unit is small and that requires parametric assumptions, FDA is a flexible non-parametric method that can handle cases in which many observations per unit are recorded. The only requirement in FDA is a sufficiently large amount of data so that the curve can be adequately approximated [9]. The main idea is to represent and analyze observations that are curves in nature rather than univariate or multivariate. This is done by first creating a functional representation of the observations using smoothing methods, such that a flexible family of functions (such as polynomials or sinusoidals) are used to fit the curves. Unlike time-series modeling, here we assume that each observation is a time series, and that the complete sample of time series forms a family. The resulting set of curves forms the “functional object”, which can then be analyzed using various

statistical methods. The functional representation itself allows data compression, since the data can be stored and manipulated in the form of basis functions and some coefficients.

There are two approaches that generalize ordinary statistical methods to the functional setting: One is to place a grid on the curves thereby obtaining multiple “slices” of the data, then to apply statistical methods pointwise, and finally to interpolate the results. For example, if we have a set of weight curves for 100 children over a period of 3 years, we could sample them at monthly intervals, compute the pointwise average weight, and then interpolate the resulting monthly means. The second approach is to apply the statistical procedure directly to the curve coefficients. This is very useful and computationally efficient when the curves are represented by a linear set of functions (e.g., B-splines). In that case, any linear operation such as computing the mean or carrying out a linear regression can be applied directly to the function coefficients!

In the next section we show why FDA is suitable for handling online auction data, and how it is advantageous over static statistical methods.

3. FDA Implementation

The process of modeling data using functional data analysis begins by representing each bid sequence by a curve. The assumption is that there is an underlying price curve which manifests itself as points over time. After the underlying curve is estimated, or “recovered”, we follow the traditional statistical process of analysis: We start by plotting the raw data (which are the curves), then we compute summaries and characterize the overall features of the dependent variable. We continue the exploration using methods such as cluster analysis for learning about natural grouping of auctions. Finally, we perform more formal parametric or non-parametric modeling, thereby integrating factors of interest and examining their effect on the response. These steps and their functional implementation are described next. In order to illustrate each method and what is gained by it we use a collection of data on 353 closed auctions that took place on eBay.com during November-December 2003. The auctioned items include a variety of popular items from several categories: children's items (e.g., books and posters), tickets for college football,

collectible items (e.g., Morgan Silver Dollars), fashion items (e.g., Gucci bags) and electronics (computers and computer accessories).

Step	FDA implementation
1. Pre-process	Estimate/recover underlying curve from time series
2. Explore curves	Plot and summarize curves, cluster curves
3. Explore curve dynamics	Plot and summarize derivatives and the relations between derivatives
4. Model curves / derivatives	Regress curves/derivatives on factors of interest (or other statistical methods, e.g. principal components analysis, discriminant analysis, principal differential analysis)
5. Predict and interpret	Relate curve dynamics to domain theory; Real-time forecasting of curve continuation

Table 2: Steps in Functional Data Analysis

3.1 Recovering the Functional Object and Data Smoothing

The first step in every functional data analysis (FDA) consists of recovering, from the observed data, the underlying functional object. This functional object can be a continuous curve, a shape, an image or an even more complicated phenomenon like the movement of an object through space and time.

Consider Figure 1 which displays the bid histories for four selected auctions on exclusive women's fashion accessories. We can see that the price, as reflected by the bid histories in each of these four auctions, follow an underlying trend which can be described by a continuous curve.

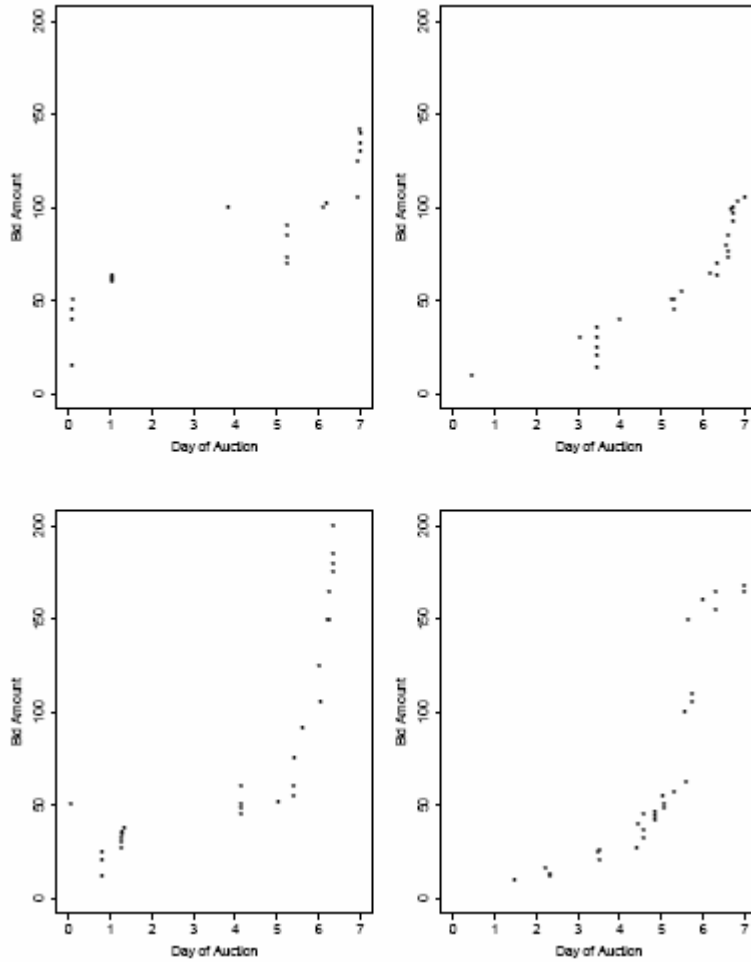


Figure 1: Bid histories for four auctions on women's fashion articles. The price increase throughout the auction exhibits different patterns across these auctions.

In the top left panel of Figure 1, for instance, the bid amount quickly increases at the beginning of the auction, then slows down over the next several days, and finally increases rapidly at the end of the auction. Interestingly, the underlying trend is not identical for the four auctions. In the bottom right panel, for instance, the bid amount only increases gradually at the beginning, then picks up speed at the end of day 5, only to slow down again near the end of the auction. Clearly, a very flexible class of models is required to accommodate these very different functional trends under one umbrella³.

There exist a variety of methods for recovering an underlying functional object from a set of data. The collection of all these methods is often referred to as *data smoothing*. For an introduction into smoothing methods see for example [31]. Here, we focus on one particularly popular method that provides a lot of modeling flexibility, the *polynomial smoothing spline*. The

polynomial smoothing spline applies the local smoothing effect of polynomials to a larger interval without the need to use high polynomial orders. Simply speaking, a spline is a piecewise polynomial function: the interval of interest is broken down into sub-intervals and within each sub-interval a polynomial is fitted. This is done in such a way that the polynomial pieces blend smoothly, so that the resulting composite function has several continuous derivatives. The edges of the sub-intervals are called *knots*. More formally, a polynomial spline of degree p can be written in the form:

$$(1) \quad f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_p t^p + \sum_{i=1 \dots L} \beta_{pi} [(t - \tau_i)_+]^p ,$$

where the constants τ_1, \dots, τ_L are a set of L knots and $u_+ = u I_{\{u \geq 0\}}$ denotes the positive part of the function u . The choices of L and p strongly influence the local variability of the function f , such that larger values of L and p result in a rougher (or more “wiggly”) f , exhibiting a larger deviation from a straight line. While this may result in a very good data fit, a very wiggly function f may not recover or identify the underlying process very well. One can measure the degree of departure from a straight line by defining a roughness penalty such as

$$(2) \quad PEN_m = \int \{D^m f(t)\}^2 dt,$$

where $D^m f$, $m=1,2,3,\dots$, denotes the m^{th} derivative of the function f . For $m=2$, for instance, PEN_2 yields the integrated squared second derivative of f which is sensitive to the curvature of the function f . One reason for the popularity of polynomial smoothing splines is the compromise that they achieve between data fit and variance reduction. Another reason is the immediate availability of the curves’ derivative functions.

Fitting a polynomial smoothing spline to the observed data y_1, \dots, y_n involves finding the coefficients $\beta_0, \beta_1, \dots, \beta_p, \beta_{p1}, \dots, \beta_{pL}$ of (1) that minimize the penalized residual sum of squares

$$(3) \quad Q_{\lambda,m} = \lambda P_m + \sum_{i=1 \dots n} \{y_i - f(t_i)\}^2 ,$$

where the smoothing parameter $\lambda \geq 0$ controls the trade-off between the data-fit, as measured by the summation on the right-hand side of (3), and the local variability of the function f , measured by the roughness penalty PEN_m in (2). Using $m=2$ in (3), for instance, leads to the commonly-encountered cubic smoothing spline, a polynomial of degree three. More generally, a smoothing spline of order m is a polynomial of degree $p=2m-1$. Figure 2 illustrates the effect of different values of m and λ on the smoothing function f .

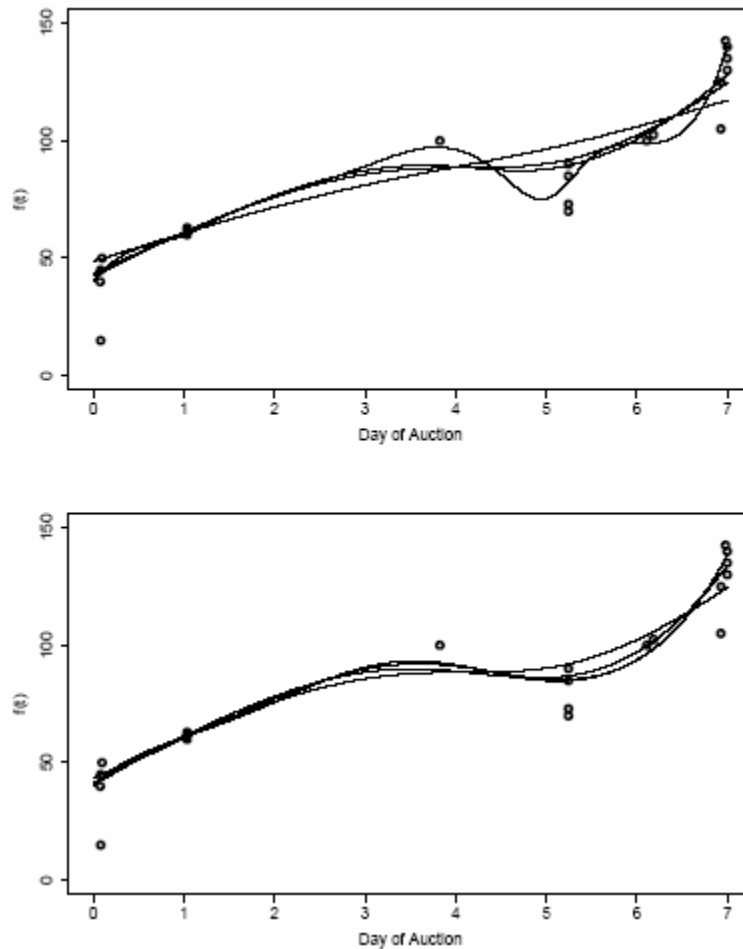


Figure 2: Different smoothing of the same data: The top panel compares smoothing splines of order $m = 2$ with values of $\lambda = 0, 0.5, 1, 10$. The spline becomes smoother as λ increases. The bottom panel compares smoothing splines of order $m = 2, 3, 4, 5$ using $\lambda = 1$. Higher order splines produce smoother curves.

Minimization of the penalized residual sum of squares (3) is done in a way very similar to the minimization of the least squares operator in standard regression analysis (see Appendix A). Many software packages exist to fit polynomial smoothing splines. In this work we use the *pspline* module. This and many more FDA functions are freely available online for R ([21]), SPlus and Matlab software (ego.psych.mcgill.ca/misc/fda/software.html).

3.2 Summarizing and Visualizing Functional Data

As in any classic statistical analysis, the first step in FDA consists of summarizing and visualizing the data (see [29] for online auction visualizations). Figure 3 shows the smoothing splines for the 353 completed eBay auctions described above. We fit the splines to logs of the bid amounts⁴ in order to capture periods of very quick increases in price, similar to growth curve models in economics or biology. We can see that there is a large amount of variability between the curves and that an overall pattern is hard to identify.

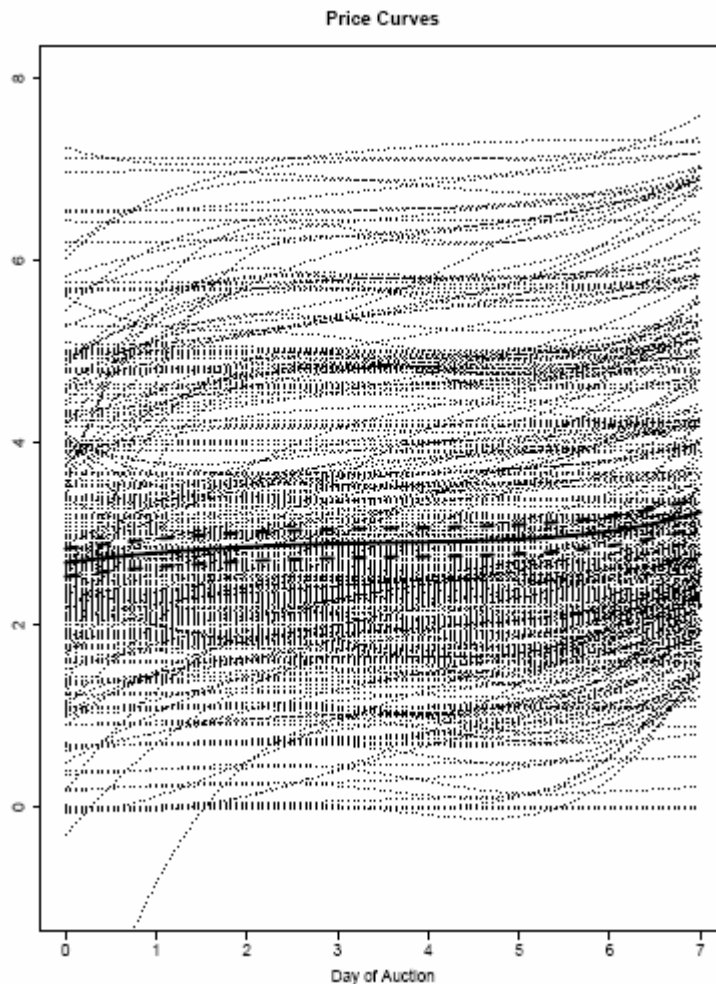


Figure 3: Summarizing functional data -- the 353 smoothed price curves with their point-wise average (thick solid line) and 95% confidence bounds (thick dashed lines).

One way to summarize the information contained in functional data is in a point-wise manner. That is, let $f_i(t)$ denote the smoothing spline pertaining to auction i . Using an evenly spaced grid of points t_1, t_2, \dots, t_G , we can sample the spline at each grid-point, leading to a set of function values $f_i(t_1), f_i(t_2), \dots, f_i(t_G)$ for each individual auction. Standard summary measures can now be applied to this grid directly. For instance, in order to determine the average trend in the functional data, the sample mean can be computed at each grid-point

$$(4) \quad \bar{x}(t_g) = \frac{1}{353} \sum_{i=1}^{353} f_i(t_g), \quad g = 1, \dots, G.$$

In a similar fashion we can compute the sample standard deviation for each grid point, leading to a set of values, say, $s(t_G)$. Using standard statistical reasoning, we can now derive point-wise confidence bounds

$$(5) \quad \bar{x}(t_g) \pm z_{\alpha/2} \cdot s(t_g) / \sqrt{353}$$

where z_α denotes the α percentile of the standard normal distribution. After interpolation, these confidence bounds can be plotted and used as a visual summary of the functional data. Figure 3, for instance, shows the point-wise mean (thick solid line) and point-wise 95% confidence bounds (thick dashed lines) for the 353 auctions. We can see that, on average, the bid amount increases gradually over the duration of the auction. Towards the end of the auction, however, the rate of increase changes.

The information in the functional data can also be summarized in other ways. Using the median instead of the mean (together with confidence bands based on, say, the 5th and 95th percentiles), for instance, is a useful alternative that is robust to outliers. A completely different approach is to visualize the *dynamics* in the observed data. This is especially useful in the auction context, because we can think of the current highest (or second highest) bid as a moving object that travels at a certain pace throughout the auction. Attributes that are typically associated with a moving object are its *velocity* (or its *speed*) as well as its *acceleration*. Given an object with a certain mass, velocity is proportional to the object's *momentum*, while acceleration is proportional to its *force*. Velocity and acceleration can be computed for each auction via the first and second derivatives of $f_i(t)$, respectively.

Figure 4 shows the velocity and acceleration for the 353 auctions. First, notice the significantly smaller amount of variability compared to Figure 3. The values of the products that are auctioned in these 353 auctions are highly variable and range from only a few dollars (e.g. children's poster of the movie *The Lord of the Rings*) to several thousand dollars (e.g. *Dell Inspiron computer*). So it is not surprising to see high variability in bid values/position (Figure 3). On the other hand, the rate at which the current bid changes from one value to the next is often much less variable. This can be observed in Figure 4, which describes the velocity and acceleration of the price in these auctions. Thus, summarizing the dynamics can be a very useful tool to learn about similarities

and differences between auctions at different levels of their dynamics. It is also useful for comparing a variety of possibly very diverse auction categories, as will be seen next.

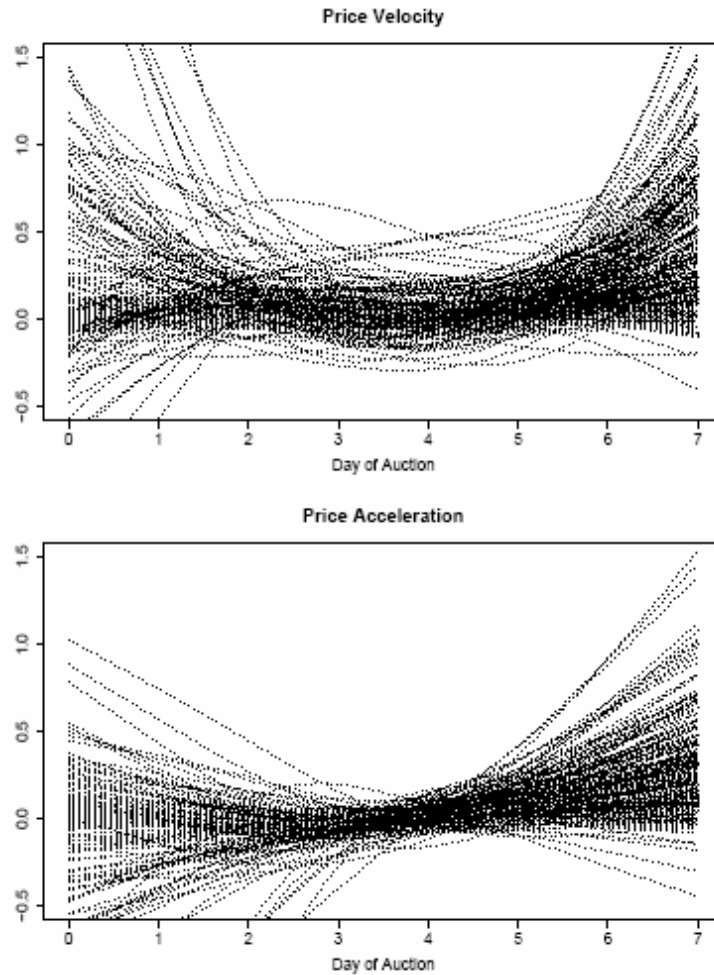


Figure 4: Dynamics of price -- the 353 price velocity (1^{st} derivative) and price acceleration (2^{nd} derivative) curves.

3.3 Functional Cluster Analysis

Exploratory statistics is typically concerned with detecting patterns within a large set of possibly high-dimensional data. In functional data analysis, since each data point consists of a continuous curve, the data is infinite-dimensional! This poses new challenges for the generalization of traditional exploratory tools like cluster analysis.

One way to handle the high dimensionality of functional data is to use a low dimensional representation of the infinite-dimensional curve. Let $\hat{\beta}_i$ denote the vector of estimated spline coefficients for the i^{th} smoothing spline. Notice that $\hat{\beta}_i$ is of finite (and typically relatively low) dimension. Moreover, within the set of all splines of order m , $\hat{\beta}_i$ determines the shape of $f_i(t)$ uniquely. Therefore, each spline (describing an auction) is exactly determined by its spline coefficients $\hat{\beta}_i$, and we use the set of coefficients as the low-dimensional representation of the infinite-dimensional spline.

Cluster analysis is a useful tool for exploring natural grouping of observations in a dataset. The functional version of cluster analysis searches for natural groupings of splines by using their coefficients. In particular, we use a variant of the well known k-means algorithm called the k-medoids algorithm, which has the advantage of robustness to outliers (For details on the functional K-medoid clustering algorithm see Appendix B). Figure 5 shows the results of clustering the spline coefficients of the 353 auctions. Two very diverse auction-clusters are recognizable: One exhibits early dynamics (left) and the other late dynamics (right). In the **early activity cluster**, price-velocity is high at the beginning of the auction. However, its acceleration is negative during that time and hence the velocity drops to almost zero towards the middle of the auction. Towards the end of the auction velocity picks up again, although this increase is rather slow. The **late activity cluster** exhibits different dynamics: In this cluster price-velocity and acceleration are near zero at the beginning of the auction. Velocity remains low until about day 5. From there on it increases sharply towards the end of the auction. The strong surge in price dynamics is even more visible in the price-acceleration where the slope continues to increase until the very end.

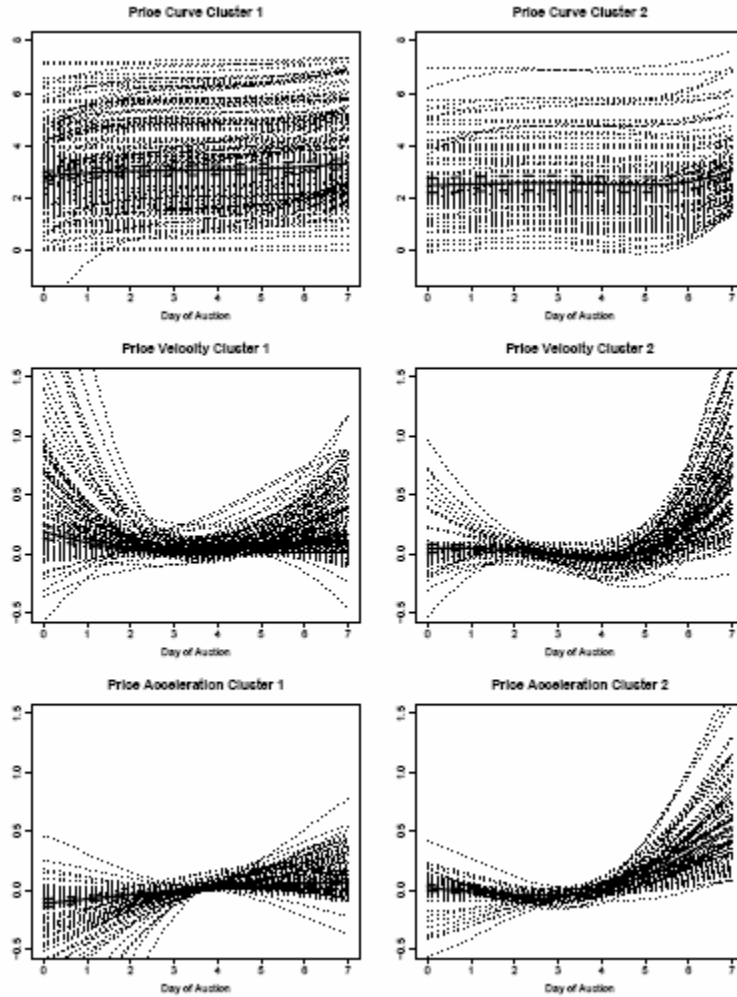


Figure 5: Cluster-specific price curves and dynamics with point-wise averages (thick solid line) and 95% confidence bounds (thick dashed lines). Cluster 1 exhibits early dynamics, while cluster 2 exhibits late dynamics.

The next step is to find what other features distinguish the clusters. Figure 6 compares the item categories between the two clusters. It can be seen that the two clusters are similar, indicating that similarities and differences in auction dynamics are global rather than category-specific.

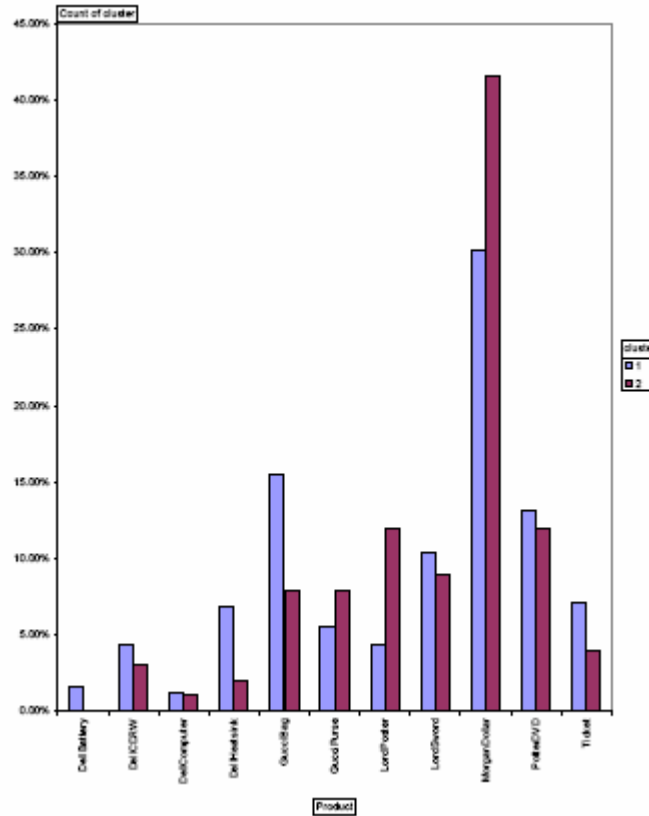


Figure 6: Cluster-specific distribution of item categories. The dataset includes auctions on items of a variety of categories and a range of prices. The two clusters are similar with respect to category distribution.

Table 3 gives the mean and standard error of different numerical variables that are recorded in the bid histories, for each of the clusters. From the five variables two seem to differ significantly from one cluster to the other: the late activity cluster has a lower average opening bid and a higher average number of bids. This is in line with evidence and theory about lower opening prices attracting more bids ([1], [28], [17]). We further examine the three-way relationship between opening bid, number of bids, and closing prices and compare the two clusters. Figure 7 shows a scatterplot of price vs. the opening bid by cluster, with circle size proportional to the number of bids. We make the following observations:

1. Most of the auctions starting at \$0.99-\$1 are in the early dynamics cluster (cluster 1). Thus, although the late dynamic cluster has a lower average opening bid, the majority of “standard” \$1 opening bid auctions fall in the early dynamic cluster. So a standard \$1 opening bid is associated with early price dynamics.

2. The early activity cluster appears to have many auctions with low-price-low-opening-bid pairs (falling on a straight line), which attracted only 1-2 bids (small circles). None of the variables in our dataset shed light on the reason for these “undesirable” items.
3. Auctions with many bids tend to have high closing prices in both clusters, as expected by the above three-way relationship. However, in the early dynamics cluster (cluster 1) the opening bid does not appear to play a role.

	(log) opening bid	(log) price	(log) seller feedback	(log) winner feedback	(log) number of bids
Early Activity Cluster	2.32 (0.113)	3.34 (0.10)	5.46 (0.12)	4.52 (0.10)	1.35 (0.07)
Late Activity Cluster	1.76 (0.123)	3.32 (0.11)	5.56 (0.15)	4.84 (0.13)	2.20 (0.05)

Table 3: Mean and (standard error) for several auction attributes

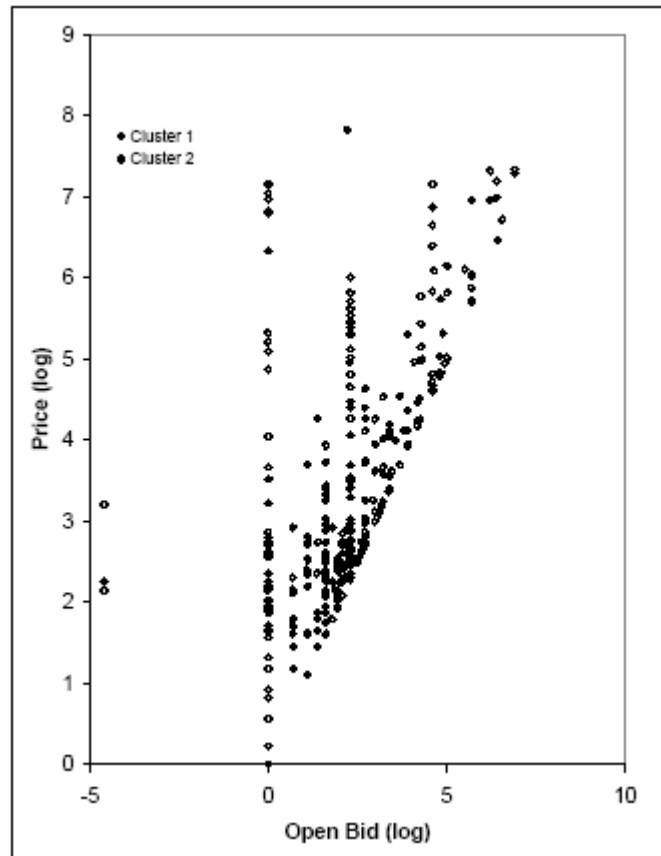


Figure 7: Price vs. opening bid, by cluster. Circle size is proportional to the number of bids.

These are a few illustrations of the types of questions and insights that can be derived by exploring the relationship between auction design, auction dynamics, and auction results. Functional cluster analysis is especially suitable for exploring dynamic hypotheses where the price dynamics are described as a flowing process that is influenced by and/or has influence over “traditional” variables such as the number of bids, the seller’s rating, and the closing price.

3.4 Functional Regression Analysis

Much of the empirical work on online auctions use regression-type models in order to find and explain the effect of different factors on a dependent variable. In the functional context we can use regression-type models to explain and predict the shape of a curve, or its dynamics, based on a set of input variables. Unlike ordinary regression models, where the response is a scalar, in functional regression the response is an object such as a curve. For instance, in the auction setting there have been multiple studies that model the final price of the auction as the response, whereas we model the complete price curve throughout the auction as the response. This means that our dependent variable can either be the splines representing the price curves, or even their derivatives, that is the price-velocity, price-acceleration, etc.

A straightforward functional implementation of regression-type models is to fit point-wise models over a grid⁵. For instance, we can take snapshots of the splines and predictors at hourly intervals and fit a model to each snapshot. The sequence of regression coefficients is then combined by interpolation. For a more formal model formulation see Appendix C.

To illustrate how functional regression can be used to investigate research questions of interest and its advantage over a static regression model, we examine the effect of the opening bid or minimum price which is set by the seller (equivalent to the “reserve price” in auction theory, [3]) on the price formation. According to auction theory, a revenue-maximizing seller should always set a reserve price that exceeds his or her value [15]. This is true as long as bidders' values are statistically independent. However, the theoretical derivations that lead to this strategy are based on a *fixed number of bidders* N , and it has been shown that the optimal minimum bid can be very

different when bidders must incur a cost to acquire information on the auctioned item [1]. This information is acquired by the bidders throughout the auction, and therefore we would learn much from examining the effect of the minimum price as information unveils to the bidders, or in other words, on the price dynamics.

The functional cluster analysis from the previous section has already highlighted the possible role of the opening bid in distinguishing between “early” and “late” activity auctions. We now continue to a more formal exploration of this relationship and its implications.

Before we fit a functional regression model, we start by fitting the standard static regression model where the (log) final auction price is regressed on the (log) opening bid. The estimated coefficient for (log) opening bid is 0.5118 (statistically significant, $p\text{-value} < 0.000$). This implies that a 1% increase in the opening bid is associated with an average increase of 0.5% in the final auction price. Let us compare this result and its usefulness to what the functional regression offers.

Figure 8 illustrates the basic idea of functional modeling. The three panels show the influence of the (log) opening bid on the (log) position of the price as well as on its velocity and acceleration. Specifically, using the smoothing spline as the response variable and the opening bid as the predictor variable, we fitted a simple linear regression model on an equally spaced grid $0 \leq t_j \leq 7$. For each of these grid-points, we obtained an estimate for the coefficient of the (log) opening bid, $\hat{\beta}_j$, together with an estimate of its standard error. The solid line in the top panel shows the interpolation of these parameter estimates while the dotted lines correspond to the resulting 95% upper and lower confidence bounds. The rightmost regression (at time $t_j=7$) is, in fact, almost equivalent to the static regression model of (log) final price on (log) opening bid, except that the static model uses the actual final prices whereas the functional regression uses the smoothed values. We repeated this procedure twice, using the price velocity and the price acceleration as the response variable. The results are shown in the middle and bottom panels of Figure 8.

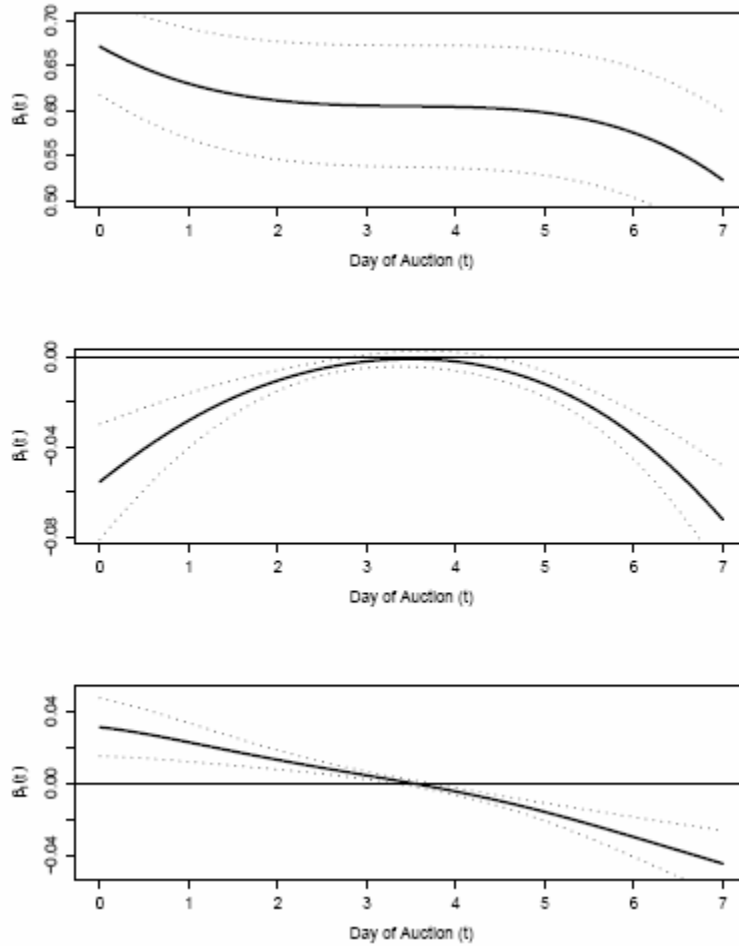


Figure 8: The estimated coefficient for opening bid in functional regression models of the form $y(t) = \beta_0(t) + \beta_1(t) \text{ Opening-bid}$. In top panel $y(t)=f(t)$ (price), middle panel $y(t)=f'(t)$ (velocity), and bottom panel $y(t)=f''(t)$ (acceleration).

The positive coefficient in the model for price indicates that the price at any point in the auction is positively associated with the opening bid, but its decreasing shape indicates the reduction in informativeness of the opening bid as the auction progresses. For price velocity the coefficient is negative and most pronounced at the start and end of the auction, indicating that during these times high opening bids are associated with reduced speed of price increase.

The top graph in Figure 8 shows that the opening bid coefficient estimate remains positive throughout the entire duration of the auction, implying a positive relationship between the opening bid and bid position. In other words, the higher the opening bid, the higher the value of the price at any time of the auction (as indicated by the static regression of closing price on opening bid). Although positive, this estimate actually declines through the auction, especially after the start and even more so towards the end of the auction. This means that the impact of the size of the opening bid on the current bid drops after the first day of the auction by

approximately 7.5%, then plateaus, and finally takes a steeper drop (~15%) as the auction approaches its end. The steep decline in the coefficient after day 5 implies that *towards the end of the auction the information contained in the opening price loses its usefulness for explaining the current price*. This is expected by auction theory due to the information acquisition cost argument: The auction start does not contain much information on the value of the auctioned item and therefore the opening bid is more crucial to bidders early in the auction, whereas once bids are placed more information is revealed on the item's valuation and the opening bid becomes less central.

To learn about the effect of opening bid on the *dynamics* of the auction we regress the derivatives of the splines on the opening bid. The middle graph in Figure 8 describes the influence of the opening bid on the price velocity. Throughout the auction the opening bid is (on average) negatively associated with the price velocity, and this relationship is statistically significant everywhere, except for a short period in the middle of the auction (between days 3-4 the 95% confidence bounds for the parameter estimate include zero, indicating that the negative association is statistically insignificant). The negative relationship means that higher opening bids are associated with slower price increases. In addition to the sign, we see that the coefficient increases during the first third of the auction, then plateaus for a day or more, and finally decreases steeply towards the end of the auction, ending with the lowest value at the end of the auction. *This implies that during the last days of the auction higher opening bids are strongly associated with slowdown in the speed of price increase, whereas lower opening bids are associated with a speeding up of the price increase*. We also learn that the closing price has the strongest relationship with the opening bid among all bids in the auction. This explains why almost every study that looked at the impact of the opening bid on the final price, using a static regression model, found a statistically significant effect.

Finally, the bottom panel in Figure 8 describes yet another dimension of the price dynamics: price acceleration. Here we see that the coefficient for the opening bid changes sign during the auction! The relationship between the opening price and price acceleration starts out positive and slowly decreases until it completely changes sign at the middle of the auction. It then continues to decrease, indicating an increasingly stronger negative relationship between the opening bid

and the price acceleration. As in the velocity case, the strongest relationship with the opening bid is at the end of the auction. The interpretation of a positive relationship is that high opening prices are associated with price acceleration, whereas lower bids are associated with deceleration. During the start of the auction, it appears that higher opening prices drive the price up by accelerating the price increase. This changes during the second half of the auction, where the negative relationship indicates that high opening bids are associated with price deceleration, and low opening prices are associated with price acceleration. In other words, the *rate of change* in current price is most sensitive to the opening price during the start and end of the auction, but in opposite ways. This could also be attributed to the two clusters in the data, and therefore a regression that includes a dummy variable separating the two clusters would give a clearer picture.

To summarize the information from the three graphs, we can say that in general *higher opening bids are associated with higher prices at any point during the auction, but this relationship weakens as the auction progresses. The dynamics of the relationship are such that high opening bids are associated with a slower increase in price compared to low opening bids, especially early in the auction and even more so towards the end of the auction. High and low opening bids also differ in the rates of change in price speed: The price increase process in auctions with high opening bids accelerates faster than in auctions with low opening bids during the first half of the auction, but slower during the second half.*

If we think of this as a car race where each car (representing an auction) has a different head start (representing the opening bid), then cars with a large head start are generally ahead of cars with little head start at any point in the race, but they go slower. The head start cars accelerate faster during the first half of the race, but then accelerate slower during the second half.

This conveys a much richer picture of the forces and dynamics that are associated with the opening bid, compared to the ordinary static regression model. Some of the dynamics were captured by the cluster analysis, which pointed out the “early” and “late” activity clusters of auctions and their relationship to the opening bid. The regression helps to quantify these phenomena and can be used to test particular hypotheses of interest.

3.5 Further Functional Methods

Other popular functional methods that extend static analyses to functional data have been developed. One method that would be useful in the context of the price curve is functional principal components analysis [24]. The idea is to condense the time dimension across price curves in order to find times during the auction that explain most of the variability across auctions. Preliminary results suggest, as expected, that the first and last day in 7-day auctions contain most of the variation across price curves. Another method that can be useful for analyzing data from experiments on online auctions (e.g., [4], [16], [22]) is functional analysis of variance (ANOVA).

There are a variety of extensions to the analysis of functional data. [33], for instance, develop a method to fit multivariate regression trees to functional data. In their approach, they consider a low-dimensional representation of the high-dimensional functional curves, either via the spline coefficients or the first several principal components of the smoothing spline. The authors then fit a standard multivariate regression tree to the low-dimensional data representation.

Approaches for non-normal data have also been extended to the functional context. [26] consider a model for binary responses that allows for the inclusion of functional covariates. Their approach is again based on a low dimension representation of the functional data which allows for an application of standard maximum likelihood methodology.

An interesting extension of FDA has been done by [23], using differential equations. The authors use functional data methodology to investigate the dynamics of the nondurable goods index. Most notably, they achieve this goal by developing and fitting differential equation models to a functional representation of the monthly production index for nondurable goods. Applying this methodology in the context of the price formation process in online auctions, we found that some types of auctions are indeed well captured by a second-order homogenous differential equation, while others do not (see [11] for details).

4 Discussion

We have shown how functional data analysis is a useful statistical approach for modeling and analyzing online auction data. It meets the criteria of being able to address research questions of interest while maintaining most of the information contained in the raw data. Furthermore, it enables to address questions about the dynamic aspect of the process and uncover such structures. In a sense it generalizes analyses that are based on a snapshot of the data at a single time point. By treating an entire curve, such as the price over the entire auction, as the dependent variable, we can investigate not only the static closing prices but also their dynamics. We can explore the curve on its own in order to characterize it, and we can explore the effect of different factors on the curve. The challenging part in FDA is the initial step of recovering the underlying curve. The smoothing step has several parameters that need to be determined: the family of functions and the degree of smoothness. Normally, the type of approximating function is based on the nature of the data: for cyclical data sinusoidal functions are useful, while for monotone functions a monotone function is more suitable, both conceptually and in many cases practically. It should be noted that different types of functions require varying computational burden. The choice of function is therefore also dependent on the number of observations and on the type of analysis to be performed. With respect to the degree of smoothness, the choice of knots and the smoothing penalty can have a significant impact on the resulting curves. There are a few criteria for performing automated parameter choices, but these can perform poorly in some circumstances. It is better practice to visually inspect the fitted curves and their fit to the actual data in order to make sure that a reasonable fit is obtained. In the auction setting an additional challenge is the sparsity of bids during mid-auction, and the presence of auctions with very few bids. Fitting splines requires a very heavy smoothness penalty in such cases, and can require an initial “light smoothing” pre-processing step. From a computational point of view, when the computations are performed on a grid of timepoints, computationally intensive methods will lead to very time consuming analyses. If the goal is real-time analysis, such methods should be substituted with simpler ones. Finally, as this method allows the analysis of a continuous process, interpreting the output requires careful attention to the time aspect, since this dimension is absent in ordinary statistical methods.

As pointed out above, a promising facet of functional data analysis is its relationship to differential equation models. In traditional auction theory many results are derived using differential equations. Our next step is to try and connect the two threads and see what relates or differentiates offline from online auction mechanisms using this approach. In particular, we would like to be able to examine auctions in the broader context of the entire auction market. One step in this direction is to borrow the concept of energy from physics and to talk about “auction energy” and even “market energy”. In common value auctions we can treat the common valuation of the item as the amount of potential energy that the auction has. This is a finite amount and at the start of the action is equally distributed over the auction duration. Within the market of auctions for such items, items that are of interest to buyers will “attract energy” from the market having a larger amount of initial energy. When an auction takes place, each placed bid consumes energy in proportion to its value or impact (higher bids consume more auction energy). Once a bid is placed the remaining energy is “re-balanced” or reallocated so that it is once again equally distributed across the remaining duration of the auction. Using the energy concept we can explain the results that we found regarding the relationship between the price formation and the opening bid: The opening bid, like any other bid, consumes auction energy. Therefore a low opening bid consumes only a little of the auction energy, leaving more potential energy untouched compared to a high opening bid. Graphing the smoothed price curve over time (e.g., Figure 3) describes the “energy consumption meter” from the auction start to its end. The price velocity (e.g., Figure 4, top) describes the level of energy consumption at each point in the auction. This is where we would expect to see bouts of consumption when a high bid is placed. The price acceleration (e.g., Figure 4, bottom) reflects the changes in energy consumption throughout the auction. Further study of the concept of auction energy is needed in order to gain insights into the forces that drive the price evolution and other related processes.

The dynamic results are useful for various managerial applications: Currently eBay allows the seller in an auction to specify a “buy-it-now” value. This value is set before the auction start and does not change. Using the dynamic model for price dynamics throughout the auction, one could design a *dynamic “buy-it-now”* feature. The dynamic value would change to reflect the change in interest and willingness-to-pay as expressed by the price curve [5].

Although our application is focused on online auction data, functional data analysis can be a valuable tool for modeling other internet and e-commerce data. The online environment in many cases involves an ongoing process (such as user interaction or market change) which can be captured by a curve. In such cases the dynamics are central to design and decision making. One example of ongoing user interactions is online product ratings. [7], who look at online book reviews and [8], who look at online movie ratings both use ratings data to investigate word-of-mouth effects and their impact on sales. In particular, [7] compare the number of book ratings, review lengths, and rating distributions of several books on Amazon vs. BN.com. Here the ratings are treated as static, in the sense that the rating evolution is not considered. [8] compare ratings on the opening weekend of the movie to the second week and to later ratings and find interesting differences. However, their actual analyses are based on static versions of the ratings. In both applications it seems beneficial to capture the entire rating evolution: An FDA approach would treat the rating evolution or the cumulative number of ratings for a single event as a curve. Curves can then be compared and their dynamics explored and modeled.

Other examples where the online process evolution can contain insightful information and where the data are readily available are assessing website usability by inspecting user browsing patterns (e.g., the number of webpages or clicks that a user goes through in order to complete an online transaction.) Another example is the dynamics in newsgroups, where we can track the number of new postings over the lifetime of a thread. A third type of process where dynamics are crucial to performance and design is sales of time-sensitive goods⁶. In the airline and hotel industries, for example, the prices of tickets or hotel rooms change over time as a function of demand and the deadline after which the goods are no longer valuable (e.g. when the flight leaves or the hotel fills up). In such cases information about the dynamics can be integrated directly into the pricing to reflect the dynamic demand over time.

Finally, with the latest advances in software agent collection methods that will supply longitudinal or repeated auction data [12], functional analysis will be especially useful because the collected data will be functional by nature. For example, the auction closing price for a certain item over time is no longer a single number, and can be described by a curve. We

anticipate that FDA will play a major role in longitudinal analyses that will follow from such data.

Appendix A: Estimation of the Smoothing Spline

To describe the minimization of the penalized residual sum of squares in (3), we define the $(L + p + 1)$ vector of spline basis functions

$$(6) \quad \mathbf{x}(t) = (1, t, t^2, \dots, t^p, [(t-\tau_1)_+]^p, \dots, [(t-\tau_1)_+]^p)$$

and notice that we may write the spline in (1) as $f(t) = \mathbf{x}(t)\boldsymbol{\beta}$, where

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p, \beta_{p1}, \dots, \beta_{pL})'$ is the $(L + p + 1)$ parameter vector. The roughness penalty in Equation (2) can now be written as

$$(7) \quad \text{PEN}_m = \boldsymbol{\beta}' \mathbf{D} \boldsymbol{\beta},$$

where the symmetric positive semi-definite penalty matrix \mathbf{D} is defined as

$$(8) \quad \mathbf{D} = \int \{ \mathbf{D}^m \mathbf{x}(t) \}' \{ \mathbf{D}^m \mathbf{x}(t) \} dt .$$

We can now rewrite the penalized residual sum of squares in (3) as

$$(9) \quad Q_{\lambda,m} = \lambda \boldsymbol{\beta}' \mathbf{D} \boldsymbol{\beta} + \sum_{i=1 \dots n} \{ y_i - \mathbf{x}(t_i) \boldsymbol{\beta} \}^2$$

Let $\mathbf{y} = (y_1, \dots, y_n)'$ denote the vector of the prices and define the matrix of spline basis functions

$$(10) \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}(t_1) \\ \mathbf{x}(t_2) \\ \vdots \\ \mathbf{x}(t_n) \end{pmatrix}$$

Equation (9) can now be rewritten as

$$(11) \quad Q_{\lambda,m} = \lambda \boldsymbol{\beta}' \mathbf{D} \boldsymbol{\beta} + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Setting the gradient of the right hand side of (11) equal to zero and rearranging terms yields the estimating equations

$$(12) \quad (\mathbf{X}'\mathbf{X} + \lambda\mathbf{D}) \boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$$

Solving for $\boldsymbol{\beta}$ in (12) gives the penalized spline estimator

$$(13) \quad \hat{\boldsymbol{\beta}}_{ps} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{D})^{-1} \mathbf{X}'\mathbf{y}$$

We note that the Hessian matrix of (11) is

$$(14) \quad 2(\mathbf{X}'\mathbf{X} + \lambda\mathbf{D}).$$

Since the matrix $\mathbf{X}'\mathbf{X}$ is positive definite and $\lambda\mathbf{D}$ is positive semi-definite, the Hessian matrix is positive definite and, hence, $\hat{\boldsymbol{\beta}}_{ps}$ in (13) indeed minimizes the penalized residual sum of squares in (11).

Appendix B: Functional K-Medoids Clustering Algorithm

For two vectors of coefficients $\hat{\beta}_i$ and $\hat{\beta}_{i'}$, let $d_{i,i'} = D(\hat{\beta}_i, \hat{\beta}_{i'})$ denote a measure of dissimilarity between them. A variety of dissimilarity measures exist. By far the most common measure is the Euclidian distance, $D(\hat{\beta}_i, \hat{\beta}_{i'}) = \|\hat{\beta}_i - \hat{\beta}_{i'}\|$. The K-medoids algorithm (e.g. [14], [10]) is an iterative procedure whose goal is to find a partition of the set of all spline coefficients that minimizes the within-cluster dissimilarity

$$(15) \quad W_K = \sum_{k=1}^K \sum_{i' \in I_k} d_{i,i'}$$

where I_k denotes the set of indices pertaining to the elements of the k^{th} cluster, $k = 1, \dots, K$. The K-medoids algorithm achieves this goal in iterative fashion, by alternating between two steps. In the first step, cluster centers are determined. That is, given a current data-partition, one finds the observation in the k^{th} cluster that minimizes the total distance to the other points in that cluster:

$$(16) \quad i_k^* = \arg \min_{i \in I_k} \sum_{i' \in I_k} d_{i,i'}$$

Then, $c_k = \hat{\beta}_{i_k^*}$, $k = 1, \dots, K$, is the current estimate of the center of cluster k . The second step reassigns observations to their nearest cluster. That is, given a current set of cluster centers $\{c_1, \dots, c_K\}$, one finds a new partition by assigning $\hat{\beta}_i$ to the cluster k for which

$$(17) \quad k = \operatorname{argmin}_{1 \leq k \leq K} D(\hat{\beta}_i, c_k)$$

These two steps are repeated until the assignments do not change any further.

Appendix C: Functional Regression Formulation

Let

$$(18) \quad \mathbf{Y}(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{pmatrix}$$

be a $n \times I$ vector of curves that we wish to model. If we wish to model the position of the price, for instance, then these curves could be given by the smoothing splines themselves, $y_i(t) = f_i(t)$. On the other hand, if the goal is to find a suitable model for the velocity, then we can put $y_i(t) = f_i'(t)$. We define a q vector of parameter curves $\boldsymbol{\beta}(t)$ as

$$(19) \quad \boldsymbol{\beta}(t) = \begin{pmatrix} \beta_1(t) \\ \beta_2(t) \\ \vdots \\ \beta_n(t) \end{pmatrix}$$

If \mathbf{Z} denotes a suitable (and known) $n \times q$ design matrix, then the functional linear model attempts to find $\boldsymbol{\beta}(t)$ such that the expected value of $\mathbf{Y}(t)$ is $\mathbf{Z}\boldsymbol{\beta}(t)$ for each value of t .

The above problem can be written in a way very similar to the least squares minimization objective of ordinary regression. Let

$$(20) \quad \text{ISSE}(\boldsymbol{\beta}) = \int \|\mathbf{Y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)\|^2 dt$$

denote the integrated residual sum of squares, where $\|\cdot\|$ denotes the Euclidian norm. The goal is to find $\boldsymbol{\beta}(t)$ that minimizes ISSE. Since there is no particular restriction on the way in which $\boldsymbol{\beta}(t)$ varies as a function of t , one can minimize ISSE by minimizing

$$(21) \quad \|\mathbf{Y}(t) - \mathbf{Z}\boldsymbol{\beta}(t)\|$$

individually for each t ([24]). In particular, the most straightforward method of finding an estimate $\hat{\beta}(t)$ is to find $\hat{\beta}(t_j)$ that minimizes (21) for a suitable grid of values t_1, t_2, \dots , and then to interpolate these values.

References

- [1]. Bajari, P. and Hortacsu, A. Cyberspace auctions and pricing issues: A review of empirical Findings, *Working Paper #02-005*, Economics Department, Stanford University (2002).
- [2]. Bajari, P. and Hortacsu, A. Winner's curse, reserve prices and endogenous entry: Empirical insights from ebay. *RAND Journal of Economics*, 34 (2), (2003), 329-355.
- [3]. Bajari, P. and Hortacsu, A. Economic insights from internet auctions. *Journal of Economic Literature*, 42 (2), (2004), 457-486.
- [4]. Bapna, R., Goes, P., and Gupta, A. Comparative analysis of multi-item auctions: Evidence from the laboratory. *Decision Support Systems*, 32, (2001), 135-153.
- [5]. Bapna, R., Goes, P., Gupta, A., and Karuga, G. Predicting bidders' willingness to pay in online multi-unit ascending auctions: Analytical and empirical insights. (2004), *Working Paper*, School of Business, University of Connecticut.
- [6]. Besse, P. C., Cardot, H., and Stephenson, D. B. Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27(4), (2000), 673-687.
- [7]. Chevalier, J. A. and Mayzlin, D. The effect of word of mouth on sales: online book reviews, (August 6, 2003), Yale SOM Working Paper No's. ES-28 & MK-15.
- [8]. Dellarocas, C. N., Awad, N. and Zhang, X. Using online reviews as a proxy of word-of-mouth for motion picture revenue forecasting, (May 10, 2004), Working paper, MIT.
- [9]. Faraway, J. J. Regression analysis for a functional response. *Technometrics*, 39,(1997), 254-261.
- [10]. Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.
- [11]. Jank, W. and Shmueli, G. Profiling price dynamics in online auctions using curve clustering, (2005), *Working paper*, Smith School of Business, University of Maryland.
- [12]. Kauffman, R. J., March, S. T., and Wood, C. A. Agent sophistication: design aspects for data-collecting agents. *International Journal of Intelligent Systems in Accounting, Finance, and Management*, forthcoming.

- [13]. Kauffman, R. J. and Wood, C. A. Revolutionary research strategies for e-business management: A philosophy of science perspective for research design and data collection in the age of the internet, (2003), MIS Research Center Working Paper #03-32, Carlson School of Management, University of Minnesota.
- [14]. Kaufman, L. and Rousseeuw, P. J. Clustering by means of medoids. In *Statistical Data Analysis Based on the L1-norm and Related Methods*, 1987, pp. 405-416.
- [15]. Krishna, V. *Auction Theory*. Academic Press, San Diego, 2002.
- [16]. List, J. A. and Lucking-Reiley, D. Bidding behavior and decision costs in field experiments. *Economic Inquiry*, 40(44), (2002), 611-619.
- [17]. Lucking-Reiley, D. Auctions on the internet: What's being auctioned and how? *Journal of industrial economics*, 48(3), (2000), 227-252.
- [18]. Ogden, R. T., Miller, C. E., Takezawa, K., and Ninomiya, S. Functional regression in crop lodging assessment with digital images. *Journal of Agricultural, Biological, and Environmental Statistics*, 7(3), (2002), 389-402.
- [19]. Onatski, A. & Kargin, S. Dynamics of Interest Rate Curve by Functional Auto-regression, *Econometric Society 2004 North American Summer Meetings* 229, (2004).
- [20]. Pfeiffer, R. M., Bura, E., Smith, A., and Rutter, J. L. Two approaches to mutation detection based on functional data. *Statistics in Medicine*, 21(22), (2002), 3447-3464.
- [21]. *R The R Project for Statistical Computing*, (2003), <http://www.r-project.org/index.html>.
- [22]. Rafaei, S. and Noy, A. Online auctions, messaging, communications and social facilitation: A simulation and experimental evidence. *Journal of Information Systems*, 11(3), (2002), 196-207.
- [23]. Ramsay, J. O. and Ramsey, J. B. Functional data analysis of the dynamics of the monthly index of nondurable goods production, *Journal of Econometrics*, 107(1-2), (2002), 327-344.
- [24]. Ramsay, J. O. and Silverman, B. W. *Functional data analysis*. Springer-Verlag, New York, 1997.
- [25]. Ramsay, J. O. and Silverman, B. W. *Applied functional data analysis: methods and case studies*. Springer-Verlag, New York, 2002.

- [26]. Ratcliff, S. J., Heller, G. Z., and Leader, L. R. Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statistics in Medicine*, 21(8), (2002), 1115-1127.
- [27]. Rossi, N., Wang, X., and Ramsay, J. O. Nonparametric item response function estimates with the EM algorithm, *Journal of Educational and Behavioral Statistics*, 27(3), (2002), 291-317.
- [28]. Roth, A. E. and Ockenfels, A. Last-minute bidding and the rules for ending second-price auctions: Theory & evidence from a natural experiment on the internet, (2000), *NBER Working Paper #7729*.
- [29]. Shmueli, G. and Jank, W. Visualizing online auctions. *Journal of Computational and Graphical Statistics*, forthcoming.
- [30]. Shmueli, G., Russo, R. P., and Jank, W. Modeling bid arrivals in online auctions, (2004), *Working paper*, Smith School of Business, University of Maryland.
- [31]. Simonoff, J. S. *Smoothing methods in statistics*. Springer-Verlag, 1996.
- [32]. Wilcox, R. T. Experts and amateurs: The role of experience in internet auctions. *Marketing Letters*, 11(4), (2000), 363-374.
- [33]. Yu, Y. and Lambert, D. Fitting trees to functional data, with an application to time of day patterns. *Journal of Computational and Graphical Statistics*, 8, (2000), 749-762.

Footnotes:

1. For recent advances and challenges in software agents development See Kauffman et al. (2004). – p2
2. For specialized visualizations for online auction data see Shmueli & Jank, 2005. – p6
3. A slight enhancement is to transform the proxy bids which can be non-monotone, to the “current price” values that form a monotonically increasing step function (see Jank & Shmueli, 2005 for further details). – p8
4. This can be applied either to the proxy bids directly or to their corresponding “current price” values, which are the values that bidders see while the auction takes place. The difference between the two is because eBay only discloses the second-highest bid at any time of the auction. – p11
5. In many cases we approximate the curves with a linear combination of B-splines. In that case we can apply linear operators (e.g., computing the mean, or fitting a linear regression model) directly to the B-spline coefficients, rather than to a grid of points. This is more computationally efficient. – p14
6. Based on private communication with Professor Michael Ball, University of Maryland. – p21