Reithinger, Florian; Jank, Wolfgang; Tutz, Gerhard; Shmueli, Galit

Working Paper

Smoothing sparse and unevenly sampled curves using semiparametric mixed models: an application to online auctions

Discussion Paper, No. 483

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Reithinger, Florian; Jank, Wolfgang; Tutz, Gerhard; Shmueli, Galit (2006) : Smoothing sparse and unevenly sampled curves using semiparametric mixed models: an application to online auctions, Discussion Paper, No. 483, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München, https://doi.org/10.5282/ubm/epub.1851

This Version is available at:
https://hdl.handle.net/10419/31032

Smoothing Sparse and Unevenly Sampled Curves using Semiparametric Mixed Models: An Application to Online Auctions

Florian Reithinger†, Wolfgang Jank‡, Gerhard Tutz†, Galit Shmueli‡

†Institut für Statistik
Ludwig-Maximilians-Universität
München

‡Department of Decision & Information Technology
Robert H. Smith School of Business
University of Maryland
College Park, MD

July 31, 2006

Abstract

Functional data analysis can be challenging when the functional objects are sampled only very sparsely and unevenly. Most approaches rely on smoothing to recover the underlying functional object from the data which can be difficult if the data is irregularly distributed. In this paper we present a new approach that can overcome this challenge. The approach is based on the ideas of mixed models. Specifically, we propose a semiparametric mixed model with boosting to recover the functional object. While the model can handle sparse and unevenly distributed data, it also results in conceptually more meaningful functional objects. In particular, we motivate our method within the framework of eBay's online auctions. Online auctions produce monotonic increasing price curves that are often correlated across two auctions. The semiparametric mixed model accounts for this correlation in a parsimonious way. It also estimates the underlying increasing trend from the data without imposing model-constraints. Our application shows that the resulting functional objects are conceptually more appealing. Moreover, when used to forecast the outcome of an online auction, our approach also results in more accurate price predictions compared to standard approaches. We illustrate our model on a set of 183 closed auctions for Palm M515 personal digital assistants.

Key words and phrases: Nonparametric methods, smoothing, mixed model, boosting, penalized splines, online auction, eBay.

1 Introduction

The technological advancements in measurement, collection, and storage of data have led to more and more complex data-structures. Examples include measurements of individuals' behavior over time, digitized 2- or 3-dimensional images of the brain, and recordings of 3- or even 4-dimensional movements of objects travelling through space and time. Such data, although recorded in a discrete fashion, are usually thought of as continuous objects represented by functional relationships. This gives rise to functional data analysis (FDA). In FDA (Ramsay and Silverman, 2002, 2005) the center of interest is a set of curves, shapes, objects, or, more generally, a set of *functional observations*. This is in contrast to classical statistics where the interest centers around a set of data vectors. In that sense, functional data is not only different from the data-structure studied in classical statistics, but it actually generalizes it. Many of these new data-structures call for new statistical methods in order to unveil the information that they carry.

Any set of functional data consists of a collection of continuous functional objects such as a set of continuous curves describing the temperature changes over the course of a year, or the price increase in an online auction. Despite their continuous nature, limitations in human perception and measurement capabilities allow us to observe these curves only at discrete time points. Thus, the first step in a typical functional data analysis is to recover, from the observed data, the underlying continuous functional object. This recovery is typically done with the help of smoothing methods.

When recovering the functional object, one encounters a variety of challenges two of which are sparse and unevenly distributed data. Smoothing methods often operate locally which means that sparse and unevenly distributed data can lead to curves that are very unrepresentative of the underlying functional object. The problem of sparse and unevenly distributed data is very acute since more and more real-world processes generate such kind of data. One example are online auctions where the arrival of data is determined by many different sources that act independently of one another, such as sellers who decides when to start and stop the auction, or bidders who decide when and where to place their bids. The situation is similar in web logs ("blogs") where the arrival of new postings depends on the arrival (and the importance) of news. Similarly, information on a patient's medical status becomes available only when the patient decides to visit a doctor. Either way, the result is irregularly spaced data which pose a challenge to traditional smoothing methods.

It is important to obtain accurate representations of the underlying functional object. Just like measurement error leads to an (unwanted) source of variation in classical statistics, poor curve representation can lead to an (additional) error source in FDA. Moreover, in FDA one often analyses *derivatives* of the functional objects in order to, say, study the dynamics of a process (Jank and Shmueli, 2005b). Then, if already the curve contains error, this error will be propagated (and magnified) to the curve-derivative. Another important area is curve-forecasting to obtain dynamic, real-time predictions of online auctions (Wang et al., 2006; Jank et al., 2006). There, if the functional object is poorly represented, then the prediction (together with the ensuing conclusions) can be far off. In this paper we propose a method that can overcome sparse and unevenly distributed data by borrowing information from neighboring functional objects. The underlying idea is very similar to that of mixed models (McCulloch and Searle, 2000).

One additional advantage of our modeling approach is that it results in conceptually more meaningful functional objects compared to previous approaches. Much of the extant literature that studies online auctions assumes independence between two auctions (Lucking-Reiley et al., 1999; Kauffman and Wood, 2003; Bapna et al., 2003; Roth and Ockenfels, 2002; Bapna et al., 2005). This assumption though is hard to justify from a practical point of view given that it is very easy for a bidder to monitor 10 or more auctions simultaneously. For instance, if a bidder participates in two auctions simultaneously, then prices in these two auctions are no longer independent of one another. Also, the independence assumption implies that two auctions for the same (or similar) item transacting during the same period of time have no affect on one another (see Jank and Shmueli, 2005a, for evidence against this assumption). This assumption is typically not made out of ignorance of the fact, but rather due to the lack of models flexible enough to account for the different types of correlation structures. Clearly, there is room for statistical thought and innovation. The method proposed in this manuscript is one attempt into that direction.

We focus here on methods that can overcome irregularly spaced data and that can also incorporate dependencies among functional objects. These methods are derived from the mixed regression model framework. In the context of regression models, much work has been done to extend the strict parametric form to include more flexible semi- and nonparametric approaches. For details see Hastie and Tibshirani (1990), Green and Silverman (1994) or Schimek (2000). For example, the P-Spline (e.g. Eilers and Marx, 1996) is very versatile and requires only an a-priori decision about a few basic smoothing parameter settings such as the location and number of knots, the order of

the spline, and the magnitude of the smoothing penalty. However, one of the disadvantages of P-Splines (and also of other smoothing methods), is the manual (or semi-manual) selection of the smoothing parameters. Another disadvantage is that they require relatively large sample sizes to produce reliable results. Furthermore, not only is the sample size important but also the sample variability. For instance, if one wishes to estimate a function over a particular region, then the results returned by P-Splines can be very poor if that region is sampled only very locally. Thus, traditional smoothing methods can be problematic if the data are sparse and only unevenly distributed. We overcome this problem using semiparametric mixed models. Our boosting approach also results in automated selection of the smoothing parameters.

This paper is organized as follows. In Section 2 we review the basics of eBay's auction mechanism and describe the data-challenges it produces. In Section 3 we describe two approaches for modeling sparse and unevenly spaced data. The first approach is the more traditional approach based on penalized smoothing splines and we demonstrate situations when it becomes unreliable. The second approach uses the ideas of mixed models. We describe the general semiparametric mixed model for estimating sparse and unevenly spaced data and describe boosting strategies to estimate the model parameters. We apply the method to a set of eBay auctions in Section 4. We conclude with final remarks in Section 5.

2 Recovering the Price-Curve in Online Auctions

In the following we motivate the problem of recovering sparse and unevenly sampled curves by considering eBay's online auctions (see `www.ebay.com`). We describe eBay's auction mechanism, the data that it generates, and the challenges involved in taking a functional approach to analyzing online auction data.

2.1 eBay's Auction Mechanism

eBay is one of the biggest and most popular online marketplaces. In 2005, eBay had 180.6 million registered users, of which over 76.8 million bid, bought, or sold an item, resulting in over 1.9 billion listings for the year. Part of its success can be attributed to the way in which items are being sold on eBay. The dominant form of sale is the auction and eBay's auction format is a variant of the second price sealed-bid auction ("Vickrey auctions", see e.g. Krishna, 2002) with "proxy bidding".

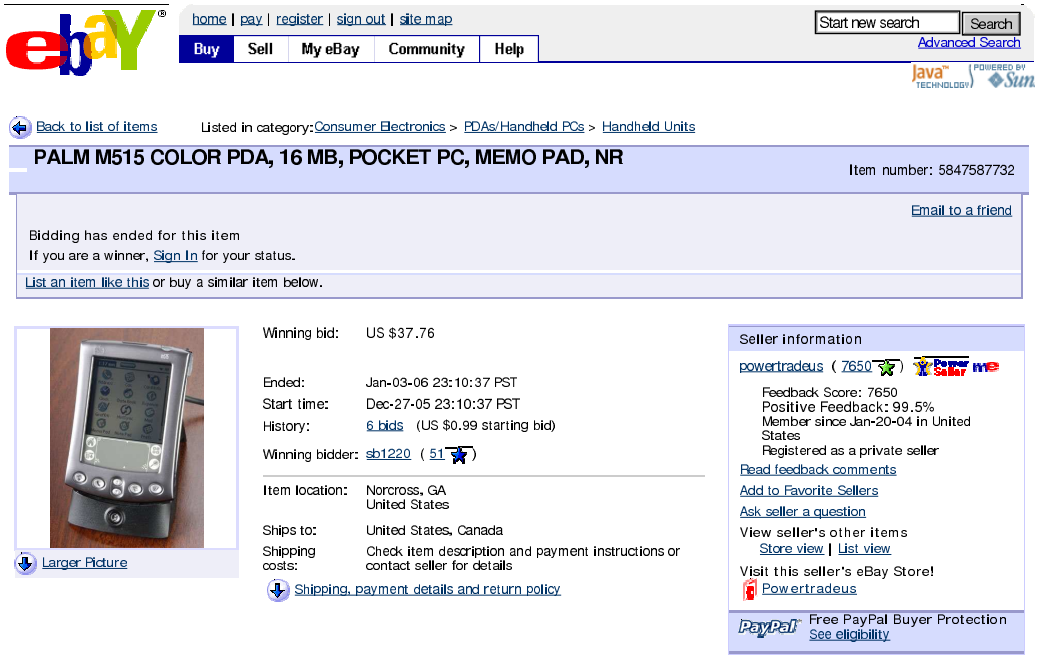
This means that individuals submit a "proxy bid", which is the maximum value they are willing to pay for the item. The auction mechanism automates the bidding process to ensure that the person with the highest proxy bid is in the lead of the auction. The winner is the highest bidder and pays the second highest bid. For example, suppose that bidder A is the first bidder to submit a proxy bid on an item with a minimum bid of $10 and a minimum bid-increment of $0.50. Suppose that bidder A places a proxy bid of $25. Then eBay's web page automatically displays A as the highest bidder, with a bid of $10. Next, suppose that bidder B enters the auction with a proxy bid of $13. eBay still displays A as the highest bidder, however it raises the displayed high-bid to $13.50, one bid increment above the second-highest bid. If another bidder submits a proxy bid above $25.50, bidder A is no longer in the lead. However, if bidder A wishes, he or she can submit a new proxy bid. This process continues until the auction ends. Unlike some other auctions, eBay has strict ending times, ranging between 1 and 10 days from the opening of the auction, as determined by the seller.

2.2 eBay's Data

eBay is a rich source of high-quality – and publicly available – bidding data. eBay posts complete bid histories of closed auctions for a duration of at least 15 days on its web site[1]. One implication of this is that eBay-data do not arrive in the traditional form of tables or spreadsheets; rather, they arrive in the form of HTML pages.

Figure 1 shows an example of eBay's auction data. The top of Figure 1 displays a summary of the auction attributes such as information about the item for sale, the seller, the opening bid, the duration of the auction, and the winner. The bottom of Figure 1 displays the bid history, that is, the temporal sequence of bids placed by the individual bidders. Figure 2 shows the scatter of these bids over the auction duration (a 7-day auction in this example). We can see that only 6 bids were placed in this auction and that most bids were placed towards the auction end, with the earlier part of the auction only receiving one bid. If we conceptualize the evolution of price as a continuous curve between the start and the end of the auction, then Figure 2 shows an example of a very sparse and unevenly sampled price-curve.

[1]See `http://listings.ebay.com/pool1/listings/list/completed.html`



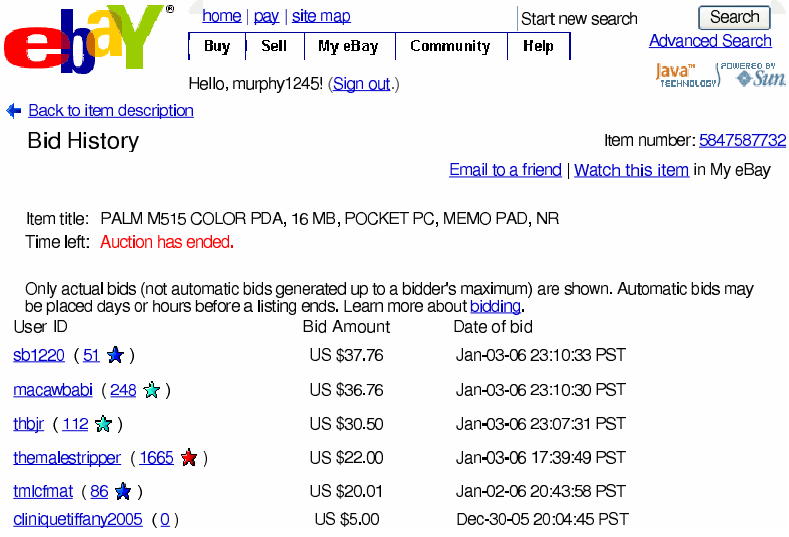


Figure 1: Bid history for a completed eBay auction. The top part displays auction attributes and includes information on the auction format, the seller and the item sold; the bottom part displays the detailed history of the bidders and their bids.

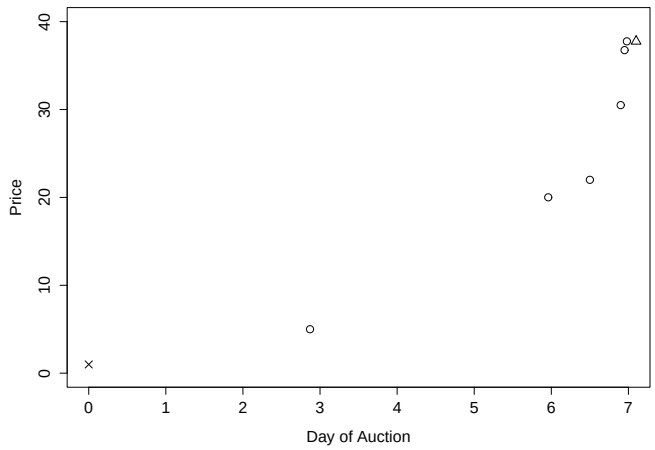


Figure 2: Scatterplot for bid history in Figure 1. The "×" marks the opening bid; the "△" marks the final price. Of the total of 6 bids, only one arrives before day 6.

2.3 Price-Curve and Data Challenges

Studying and modeling the price-curve can help in finding answers to questions such as "Does price in a typical online auctions increase sharply at first and then level-off towards the end?" Or, conversely, "Does price remain low throughout most of the auction only to experience sharp increases at the end?" And if so, "Is this price pattern the same for auctions of all types? Or do patterns differ between different product categories?" Jank and Shmueli (2005b) show that answers to these questions can help in characterizing auction dynamics and lead to more informed bidding or selling decisions. Wang et al. (2006) build upon these ideas to develop a dynamic forecasting system for live auctions (see also Jank et al., 2006). In related work, Shmueli et al. (2006) develop an interactive visualization and forecasting tool for online auction data.

One way of modeling the price-curve is via functional models. However, this modeling task is complicated due to the data structure found in online auctions. Consider again the example in Figure 2. The first step in functional data analysis is to recover, from the observed bids, the continuous price-curve. Notice, however, that only 6 bids are observed, most of them occurring at the auction end. Using traditional smoothing methods to recover a continuous curve from only 6 data points of which 5 are located at the end is not particularly meaningful or reasonable and will not lead to very representative estimates of the price-curve.

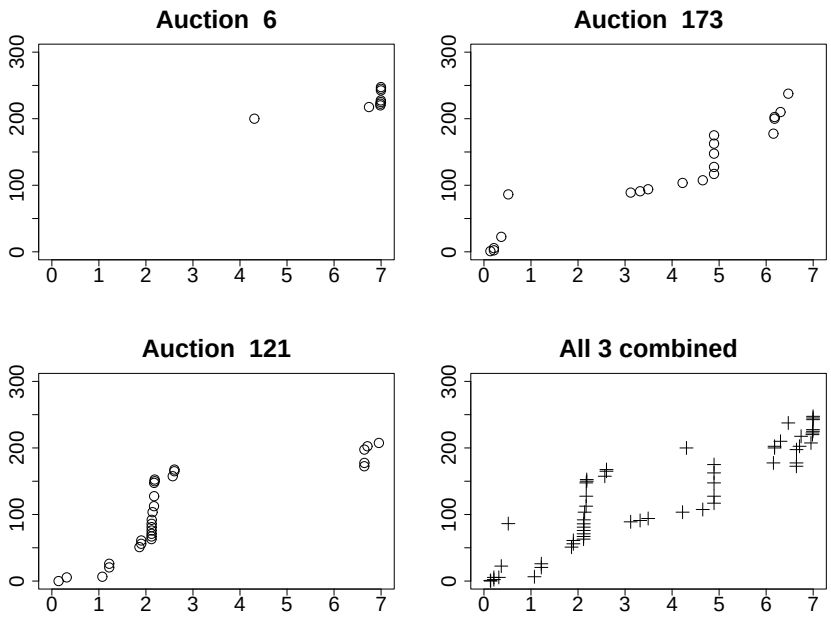


Figure 3: Three individual bid histories and their combined bids (bottom right panel).

3 Modeling Sparse and Unevenly Sampled Data

One solution is to borrow information from other, similar auctions. Figure 3 shows the bid histories for three similar auctions for the same item, labeled #6, #121 and #173. We can see that the price curve in auction #6 is only sampled at the end. Conversely, in auction #121 the price is sampled predominantly at the beginning, with no information from the middle of the auction. And finally, auction #173 contains lots of price information from the auction middle but only little from its start and end. While every auction by itself contains only partial information about the entire price-curve, if we combine the information from all three auctions, we obtain a more complete picture. This is shown in the bottom right panel of Figure 3. The idea of semiparametric mixed model smoothing is now as follows: whenever an individual auction contains incomplete information, we borrow from the combined information of all similar auctions. We describe this method more formally next.

3.1 Penalized Splines: The Challenge

The basic model for price that we consider has the form

$$\text{Price}_i(\text{t}) = \alpha_{i0} + \alpha_{(i)}(\text{t}) + \epsilon_i(t), \tag{1}$$

where t stands for time, Price$_i$(t) denotes the price of the i-th auction at time t, α_{i0} is the intercept, $\alpha_{(i)}$(t) denotes a suitable function of time, centered around zero, and $\epsilon_i(t)$ is a noise process with zero mean and variance σ_ϵ^2.

A common approach for obtaining estimates of $\alpha_{(i)}$(t) is via basis function expansion. One of the simplest basis function, the truncated power series basis of degree d, yields

$$\alpha_{(i)}(t) = \gamma_0^{(i)} + \gamma_1^{(i)} t + \ldots \gamma_d^{(i)} t^d + \sum_{s=1}^{M} \alpha_s^{(i)} (t - k_s)_+^d \ ,$$

where $k_1 < \ldots < k_M$ are distinct knots, and $\gamma_j^{(i)}$ and $\alpha_s^{(i)}$ are parameters to be estimated from the data. More generally, one uses a function of the form

$$\alpha_{(i)}(t) = \sum_{m=1}^{M} \alpha_m^{(i)} \phi_m^{(i)}(t) = \boldsymbol{\phi}_i^T(t) \boldsymbol{\alpha}_i \tag{2}$$

where $\phi_m^{(i)}$ denotes the m-th basis function, $\boldsymbol{\phi}_i^T(t) = (\phi_1^{(i)}(t), \ldots, \phi_M^{(i)}(t))$, and $\boldsymbol{\alpha}_i^T = (\alpha_1^{(i)}, \ldots, \alpha_M^{(i)})$ are unknown parameters.

Let the data be given by the pairs $(y_{is}, t_{is}), i = 1, \ldots, n, s = 1, \ldots, S_i$, where y_{is} is the price of auction i at bid number s which occurs at time t_{is}. The number of bids and their timing varies across auctions. The additive model we consider has the general form

$$y_{is} = \alpha_{i0} + \alpha_{(i)}(t_{is}) + \epsilon_{is}, \quad i = 1, \ldots, n, \quad s = 1, \ldots, S_i \tag{3}$$

where $E(\epsilon_{is}) = 0$, $\mathrm{Var}(\epsilon_{is}) = \sigma_\epsilon^2$. The above approach models each auction separately, resulting in n different function-estimates $\hat{\alpha}_{(i)}(.)$ and n different parameter-estimates $\hat{\alpha}_{i0}, i \in 1, \ldots, n$. For semi- and nonparametric regression models, Marx and Eilers (1998) propose the numerically more stable B-splines which have also been used by Hastie and Tibshirani (2000) and Wood (2004). For further properties of basis functions see also Wand (2000) and Ruppert and Carroll (1999).

Using (2) and writing $\mathbf{x}_{is}^T = [1, \boldsymbol{\phi}^T(t_{is})]$ and $\boldsymbol{\delta}_i^T = (\alpha_{i0}, \boldsymbol{\alpha}_i^T)$, (3) becomes

$$y_{is} = \alpha_{i0} + \boldsymbol{\phi}(t_{is})^T \boldsymbol{\alpha}_i + \epsilon_{is} = \mathbf{x}_{is}^T \delta_i + \epsilon_{is}, \tag{4}$$

or in matrix form

$$\mathbf{y}_i = \alpha_{i0} + \boldsymbol{\Phi}_i^T \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i = \mathbf{X}_i \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i, \tag{5}$$

with $\mathbf{y}_i^T = (y_{i1}, \ldots, y_{iS_i})$, $\boldsymbol{\Phi}_i$ has rows $\boldsymbol{\phi}_i^T(t_{is})$, $\boldsymbol{\epsilon}_i^T = (\epsilon_{i1}, \ldots, \epsilon_{iS_i})$ and $\mathbf{X}_i = [1, \boldsymbol{\Phi}_i]$. Estimates for $\boldsymbol{\alpha}_i$ may be based on the *penalized log-likelihood* for auction i

$$l_p^{(i)}(\boldsymbol{\delta}_i) = -\frac{1}{2\sigma_\epsilon^2} (\mathbf{y}_i - \mathbf{X}_i \delta_i)^T (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\delta}_i) - \frac{1}{2} \lambda \boldsymbol{\delta}_i^T \mathbf{K} \boldsymbol{\delta}_{(i)}, \tag{6}$$

where $\lambda \boldsymbol{\delta}_i^T \mathbf{K}_i \boldsymbol{\delta}_i$ is a penalty term which penalized the coefficients $\boldsymbol{\alpha}_i$. For the truncated power series an appropriate penalty is given by

$$\mathbf{K} = diag(0, \mathbf{I})$$

where $\mathbf{I}$ denotes the identity matrix and λ determines the smoothness of the function $\boldsymbol{\alpha}$. For $\lambda \to \infty$, a polynomial of degree d is fitted. P-splines use $\mathbf{K}_i = \mathbf{D}_i^T \mathbf{D}_i$ where $\mathbf{D}_i$ is a matrix of the difference between adjacent parameters yielding the penalty $\lambda \boldsymbol{\alpha}_i^T \mathbf{K}_i \boldsymbol{\alpha}_i$.

From the derivative of $l_p^{(i)}(\boldsymbol{\delta}_i)$ one obtains the estimation equation $\partial l_p^{(i)}(\boldsymbol{\delta}_i)/\partial \boldsymbol{\delta}_i = 0$ which yields

$$\hat{\boldsymbol{\delta}}_i = (\frac{1}{\sigma_\epsilon^2} \mathbf{X}_i^T \mathbf{X}_i + \lambda \mathbf{K}_i))^{-1} \frac{1}{\sigma_\epsilon^2} \mathbf{X}_i^T \mathbf{y}_i.$$

The tuning parameter lambda may be optimized by using the generalized cross-validation (GCV) criterion described in Wood (2000).

Figure 4 illustrates the performance of the penalized smoothing spline for four sample auctions. For each auction, we investigate four different smoothing scenarios: a low order (grey line) vs. a high order (black line) smoothing spline (i.e. 2nd order vs. 4th order), coupled with a low (solid line) vs. a high (dashed line) smoothing parameter ($\lambda = 0.1$ vs. $\lambda = 1$). We chose four very representative auctions out of the set of all 183 auctions.

We can see in Figure 4 that for auction #51 (upper left panel) all four smoothers deliver very similar curves, which all approximate the observed data very well. In contrast, for auction #121 (bottom left panel), the performance of the four smoothers differs greatly, especially in the middle of the auction (between days 3 and 6) where no observations are available. Moreover, notice that the higher order smoother (coupled with the lower penalty term) results in a locally variable curve where the curve increases up to day 3 but then decreases to day 6. This curve-decrease is hard to justify from a conceptual point of view since auction prices, by nature of the ascending auction mechanism, should be monotonically increasing. In fact, notice that the actual observations do increase over the same time period. However, the sparsity of the data between day 3 and day 6 causes the high-order/low-penalty smoothing spline to exhibit too much local variability. Consequently, it appears as if the lower order spline (together with the higher penalty term) is the better choice for auction #121, at least from a conceptual viewpoint. Now consider auction #141 (top right panel) which has a strong price surge at the auction-end. While the price changes only little throughout most of the auction, it jumps dramatically during the last day. Not surprisingly, only the smoother

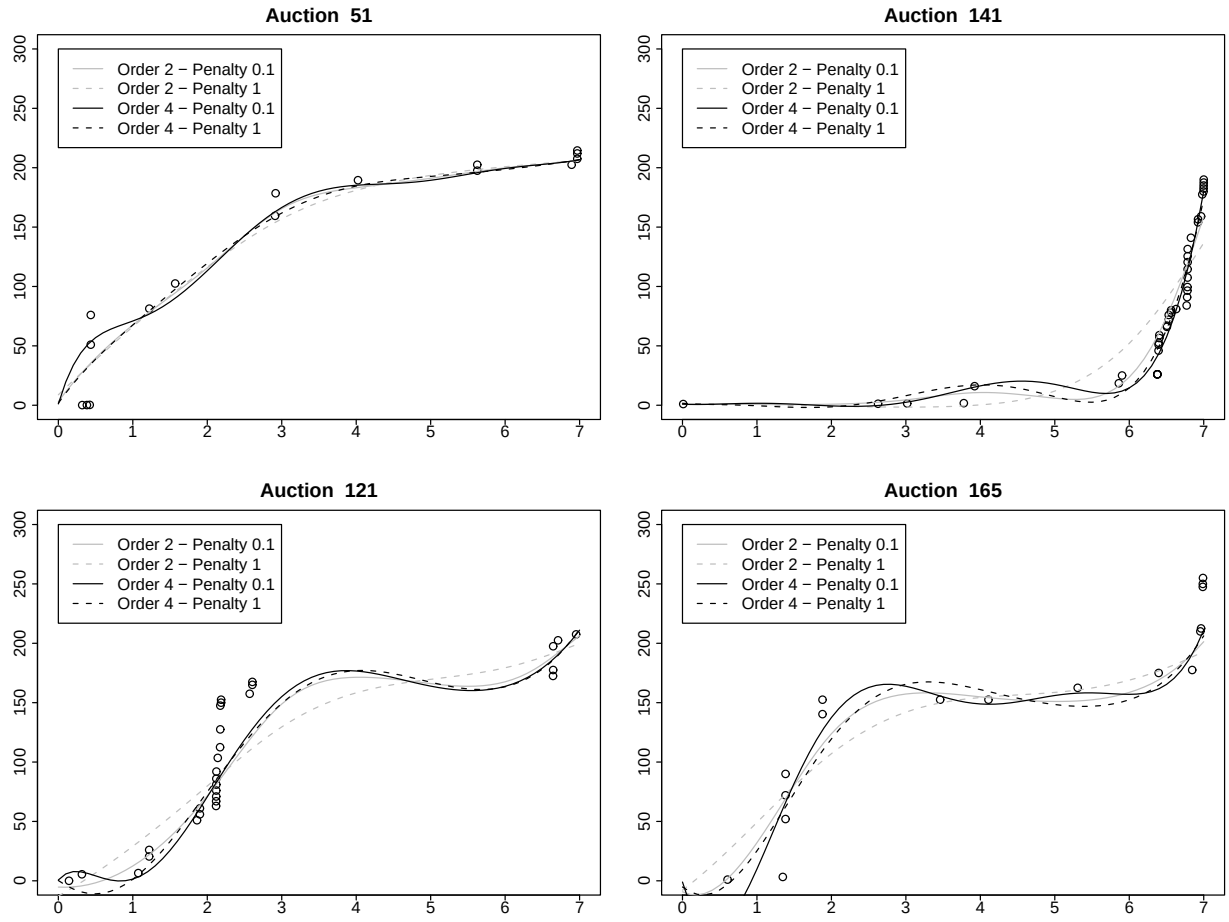


Figure 4: Performance of penalized splines using different smoothing parameters. The circles correspond to the actual live bids observed during the auction.

with the highest flexibility (order 4 and $\lambda = 0.1$) manages to capture this last moment surge in price-activity well. While the other smoothers produce a reasonable approximation for most of the auction duration, they all fail at the auction-end due to the high bidding-intensity. And finally, consider auction #165 (bottom right panel) . Interestingly, for this auction all four smoothers vary quite significantly in their fit and, more importantly, none captures the price-activity at the last moments of the auction.

In summary, although in some cases smoothing splines can produce very reasonable functional objects regardless of the smoothing parameters, in other cases the choice of the parameters can have a significant impact. In particular, while some data-scenarios call for smoother objects of lower order and higher penalty term, other scenarios require more flexible objects of higher order.

And yet, the data challenges presented in online auctions are so vast that even four different smoothers are not sufficient for accounting for all scenarios as seen in auction #165 above. There exist alternative types of smoothers that may offer some relief such as monotone smoothing splines (Ramsay, 1998); however, they can be more expensive to compute. Moreover, we find monotonicity constraints not necessary when using a mixed-model approach. We describe that approach (and our findings) next.

3.2 A Solution: Semiparametric Mixed Models

Mixed-models have been around in the statistics literature for quite a while. Yet, to date, they have found only little use in the context of functional data analysis. The basic idea of mixed effect models (or random effect models) is the presence of several data-clusters and repeat observations within each cluster (see e.g. Henderson (1953), Laird and Ware (1982) and Harville (1977)). Overviews including more recent work can be found in Verbeke and Molenberghs (2001) and McCulloch and Searle (2001). Concepts for estimating semiparametric mixed models with an implicit estimation of the smoothing parameters are described in Verbyla et al. (1999), Parise et al. (2001), Lin and Zhang (1999), Brumback and Rice (1998), Zhang et al. (1998), and Wand (2003). Bayesian approaches have been considered by e.g. Fahrmeir and Lang (2001). A different concept is the use of boosting techniques. Boosting allows fitting of additive models with many covariates. One of the major advantages of boosting is the automated selection of the smoothing parameters. Moreover, boosting techniques may be used to incorporate subject-specific variation of smooth influence functions by specifying "random slopes" on smooth effects. This results in flexible semiparametric mixed models which are appropriate in cases where a simple random intercept is unable to capture the variation of effects across subjects.

Recall that in (1), we model each auction separately, assuming independence across all n auctions. A more parsimonious (and conceptually more appealing) approach is based on semiparametric mixed model methodology. Assume that the price-curve is modeled as

$$\text{Price}_i(\text{t}) = \alpha_0 + \alpha(\text{t}) + b_{i0} + \epsilon_i(t), \tag{7}$$

where b_{i0} is a random effect with $b_{i0} \sim N(0, \sigma_b^2)$, σ_b^2 is the variance for the random intercept b_{i0} and α_0 is the (fixed) intercept for the model. Notice that in (7) we assume a common slope function $\alpha(\text{t})$ for all auctions. We also assume that the intercepts of all auctions vary randomly with mean

α_0 and variance σ_b^2. The residual error ϵ_i is assumed to be $N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{R})$, where $\mathbf{R}$ is a known residual structure, e.g. autoregressive first order. In that sense, all auctions are *conditionally* independent given the level (the random intercept). In this fashion, we can model all auctions within one parsimonious model; yet, we obtain a price-curve estimate for each auction individually.

We can obtain further modeling flexibility by also assuming random slope functions. To that end, we extend the model using flexible splines of the form

$$\text{Price}_i(\text{t}) = \alpha_0 + \alpha(\text{t}) + b_{i0} + b_{i1}\alpha(\text{t}) + \epsilon_i(t) \tag{8}$$

where b_{i0}, b_{i1} are again random effects with $\mathbf{b} := (b_{i0}, b_{i1}) \sim N(\mathbf{0}, \mathbf{Q})$. The error ϵ_i is assumed to be $N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{R})$. The model implies a common intercept and slope function for all auctions. Individual heterogeneity is induced by an auction-specific random intercept b_{i0} and an auction-specific random "slope". The covariance $\mathbf{Q}$ can be parameterized by $\boldsymbol{\rho}$, a vector of parameters to be optimized. For an unstructured covariance matrix with elements q_{11}, q_{21} and q_{22},

$$\mathbf{Q}(\boldsymbol{\rho}) = \begin{bmatrix} q_{11} & q_{21} \\ q_{21} & q_{22} \end{bmatrix},$$

the vector $\boldsymbol{\rho}$ is then the set of the elements in the lower triangular matrix of the Cholesky root $\mathbf{Q}^{1/2}$. In a more technical sense, $\boldsymbol{\rho}$ is the symmetric diagonal operator of $\mathbf{Q}^{1/2}$.

What we obtain via the model in (8) is an estimated price-curve for auction i that is characterized by the level b_{i0}, the common slope $\alpha(t)$ and the auction-specific modification $b_{i1}\alpha(t)$. Model (8) can be regarded as a restricted version of (1) using the information of the other auctions in the form

$$\alpha_{(i)}(t) \approx b_{i0} + \alpha(t) + b_{i1}\alpha(t)$$

In the following, we describe an estimation approach via boosting. Boosting allows us to jointly estimate not only the model parameters, but also the smoothing terms. Moreover, boosting as a stepwise procedure allows us to include multiplicative effects, which is not possible using REML.

3.3 Boosting and mixed model approach

Boosting originates in the machine learning community where it has been proposed as a technique for improving classification procedures by combining estimates with reweighted observations. Since it has been shown that reweighing corresponds to minimizing iteratively a loss function (Breiman

(1999), Friedman (2001)), boosting has been extended to regression problems in an L_2-estimation framework by Bühlmann and Yu (2003). In the following, boosting is used to obtain estimates for the semiparametric mixed model. Instead of using REML estimates for the choice of smoothing parameters (see Wand (2000) and Ruppert et al (2003)), the estimates of the smooth components are obtained by using "weak learners" iteratively. For observations $(y_{it}, t_{is}), i = 1, \ldots, n,\ s = 1, \ldots S_i$, one writes

$$\begin{aligned} \mathbf{y}_{is} &= \alpha_0 + \boldsymbol{\phi}(t_{is})^T\boldsymbol{\alpha} + b_{i0} + b_{i1}\boldsymbol{\phi}(t_{is})^T\boldsymbol{\alpha} + \epsilon_{is}, \quad \text{or in familiar mixed-model matrix form,} \\ \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\delta} + \mathbf{Z}_i\mathbf{b} + \boldsymbol{\epsilon}_i, \end{aligned}$$

where

$$\begin{pmatrix} \mathbf{b} \\ \boldsymbol{\epsilon}_i \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{Q}(\rho) & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2\mathbf{R} \end{pmatrix}\right),$$

and where we write $\mathbf{X}_i = [\mathbf{1},\ \boldsymbol{\Phi}_i], \boldsymbol{\delta}^T = (\alpha_0, \boldsymbol{\alpha})$ and $\mathbf{Z}_i = [\mathbf{1},\ \boldsymbol{\Phi_i\alpha}]$. Let $\mathbf{V}_i = \mathbf{V}_i(\sigma_\epsilon^2, \boldsymbol{\rho})$ denote the covariance matrix of the marginal model $\mathbf{V}_i = \mathbf{Z}_i\mathbf{Q}(\boldsymbol{\rho})\mathbf{Z}_i^T + \sigma_\epsilon^2\mathbf{R}$. Penalizing $(\alpha_0, \boldsymbol{\alpha})$ by $\boldsymbol{\delta}$ is based on the penalty matrix which for the truncated power series has the form $\mathbf{K} = \text{Diag}(0, \lambda\mathbf{I})$.

The weak learner for $\boldsymbol{\delta}$ is based on an initially fixed and very large smoothing parameter λ. By iteratively fitting of the residuals, the procedure adapts automatically to the possibly varying smoothness of the individual components. The algorithm is initialized by using an appropriate weak learner. The basic concept in boosting is that in one step the refitting of $\alpha(t_{is})$ is done by using a weak learner which in our case corresponds to large and fixed λ in the penalization term.

The algorithm works in the following way. Let $\boldsymbol{\eta}_i^{(l-1)}$ denote the estimate from the previous step. Then the refitting of residuals (without selection) is done by fitting the model

$$\mathbf{y}_i - \boldsymbol{\eta}_i^{(l-1)} \sim N(\boldsymbol{\eta}_i,\ \mathbf{V}_i(\boldsymbol{\theta}))$$

with

$$\boldsymbol{\eta}_i = \mathbf{1}\alpha_0 + \boldsymbol{\Phi}_i\boldsymbol{\alpha} + (1,\ \boldsymbol{\Phi}_i\hat{\boldsymbol{\alpha}}^{(l-1)}) \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \tag{9}$$

where α_0, $\boldsymbol{\alpha}$ are the parameters to be estimated and $\hat{\boldsymbol{\alpha}}^{(l-1)}$ is known from the previous step. Using the resulting estimates $\hat{\alpha}_0$, $\hat{\boldsymbol{\alpha}}$, the next update takes the form

$$\hat{\boldsymbol{\alpha}}^{(l)} = \hat{\boldsymbol{\alpha}}^{(l-1)} + \hat{\boldsymbol{\alpha}} \quad , \quad \hat{\alpha}_0^{(l)} = \hat{\alpha}_0^{(l-1)} + \hat{\alpha}_0.$$

The algorithm is stopped if enough complexity is in the model. Since boosting is an iterative way of fitting data the complexity of the model increases from step to step. In the beginning one fits a very robust model which adapts stepwise to the data. The complexity of the model is measured via the BIC-Criterion. Therefore in every boosting step the projection matrix of y_i on the new estimates $\boldsymbol{\alpha}^{(l)}$ and $\alpha_0^{(l)}$ is computed. Then, the trace of this matrix is used to compute the BIC-Criterion in the l-th step by setting $\text{BIC}^{(l)} = -2*\text{l}(\hat{\alpha}_0^{(l)}, \hat{\boldsymbol{\alpha}}^{(l)}) + \log(n)*\text{df}$, where *df* is the trace of the projection matrix, $\text{l}(\alpha_0^{(l)}, \hat{\boldsymbol{\alpha}}^{(l)})$ is the log-likelihood in the l-th step and n is the number of different auctions in the dataset.

The basic idea behind the refitting is that forward iterative fitting procedures like boosting are weak learners. In that sense, the previous estimate is always considered known in the last term of (9). Of course, in every step the variance components corresponding to (b_{i0}, b_{i1}) have to be re-estimated. For the complete algorithmic detail see Appendix A.

4 Application to eBay's Price Evolution

4.1 Data Description

Our data consist of 183 closed auctions for Palm M515 personal digital assistants (PDAs) that took place between March 14 and May 25 of 2003. In an effort to reduce as many external sources of variability as possible, we included data only on 7-day auctions, transacted in US Dollars, for completely new (not used) items with no added-on features, and where the seller did not set a secret reserve price. These data are publicly available at `http://www.smith.umd.edu/ceme/statistics/`.

The data for each auction include its opening price, closing price, and the entire series of bids (bid-amounts and time-stamps) that were placed during the auction. This information is found in the bid history, as shown in Figure 1.

Note that the series of bids that appear in the bid history are not the actual price shown by eBay during the live-auction; rather, they are the proxy bids placed by individual bidders (which become available only after the auction closes). eBay uses a second-price mechanism, where the highest bidder wins but pays only the second highest bid. Therefore, at each point in time, the price displayed during the live-auction is the second highest bid. For this reason, we converted the bids into "current price" values that capture the evolution of price during the live-auction. Notice that the current price data are indeed monotone increasing. This adds the extra requirement on our

smoothing method that the recovered functional object be monotone. Only few standard smoothing methods meet this requirement (Ramsay, 1998); moreover, monotonicity constraints typically also increase the computational complexity of the smoother. In the following we show that our approach, without explicitly adding any such constraints, automatically estimates the underlying monotonicity from the pooled data and imposes it on each auction-estimate individually.

4.2 Model Fit

We fit the following mixed effects model to all 183 auctions

$$s(Price_{is}) = \alpha_0 + \alpha(t_{is}) + b_{i0} + b_{i1}\alpha(t_{is}) + \epsilon_{is}.$$

Notice that α_0 and $\alpha(t)$ denote again the intercept and slope function, common across all auctions. The random effects b_{i0} and b_{i1} capture auction-individual variation. We estimate the parameters $\alpha_0, \alpha(t), q_{11}, q_{21}, q_{22}, \sigma_\epsilon^2$ using the algorithm "BoostMixed" outlined in the appendix.

Figure 5 shows the resulting curve-estimates for the first 36 auctions. The solid lines correspond to the mixed model fit; the dashed lines correspond to the ordinary penalized smoothing spline fit. We can see that the penalized smoothing splines can result in poor curve representations: in some auctions there is a lack of curvature (e.g. #3, #12), while in others there is excess curvature (e.g. #23, #34); yet in other auctions they do not produce any estimates at all due to data-sparseness (e.g. #16, #20), and yet in other auctions data-unevenness may result in very unrepresentative curves (e.g. #25, #23). Moreover, many of the curves produced by penalized splines are unsatisfactory from a conceptual point of view: for instance, in auction # 4 or #34, penalized splines result in an estimated price-path that is not strictly monotonic increasing, which violates the assumption underlying ascending auction formats.

This is very different for the estimates produced by the mixed model approach. Consider Figure **??** which shows the estimate for $\alpha(t)$ which denotes the mean slope function, common to all 183 auctions. Notice that the mean slope is monotonically increasing, as expected from an ascending auction. Moreover, the slope is steepest at the beginning and at the end of the auction which is consistent with the phenomena of early bidding and bid sniping observed in the online auction literature (Bapna et al., 2003; Shmueli et al., 2004). Mixed model smoothing takes the mean slope as blueprint for all auctions and allows for variation from the mean through the random effects b. Indeed, while the solid lines in Figure 5 all resemble the mean slope, they differ in steepness and

the timing of early and late bidding. These differences from the mean are driven by the amount (and distribution) of the observed data. For instance, auction #10 has a considerable number of bids which are distributed evenly across the entire auction-length. As a result, the estimated curve is quite different from the mean slope function. On the other hand, auction #20 only has one observation. While penalized smoothers break down with only so little information available (and do not produce any curve-estimates at all), the mixed model approach is still able to produce a reliable (and conceptually meaningful) result by borrowing information from the mean slope. It is also interesting to note that all of the price curves created by the mixed model approach are monotonically increasing. This is intriguing since, unlike the monotone smoothing splines (Ramsay, 1998), the mixed model has no built-in feature that forces the estimates to be monotone. Instead, it "learns" this feature from the pooled data which makes this a very flexible and powerful approach, suitable for many different data-scenarios.

Table 1: Estimated covariance matrix $Q(\hat{\rho})$ for the random intercept b_0 and slope b_1. The correlation is given in brackets.

	b_0	b_1
b_0	4.536 (1)	-0.619 (-0.847)
b_1	-0.619 (-0.847)	0.117 (1)

4.3 Forecasting with the Mixed Model

Another way to evaluate the quality of a smoother is via its ability to forecast the continuation of the curve. Specifically in the auction setting, we are interested in how well the estimated price curve can predict the final price of an auction. Price predictions for online auctions are becoming an increasingly important topic (Wang et al., 2006; Jank et al., 2006; Ghani, 2005; Ghani and Simmons, 2004). On eBay, an identical (or near-identical) product is often sold in numerous, often simultaneous auctions. For instance, a simple search under the key words "iPod shuffle 512MB MP3 player" returns over 300 hits for auctions that close within the next 7 days. A more general search under the less restrictive key words "iPod MP3 player" returns over 3,000 hits. Clearly, it would be challenging, even for a very dedicated eBay user, to make a purchasing decision that takes into account all of these 3,000 auctions. The decision making process can be supported via price

forecasts. Given a method to predict the outcome of an auction ahead of time, one could create an auction-ranking (from lowest to highest predicted price) and select only those auctions for further inspection with the lowest predicted price. In the following we investigate the ability of the mixed model approach to predict the final price of an auction.

We do this in the following way. We split each auction into a training set and a validation set. Specifically, we assume that the first two-thirds of the auction are observed; we estimate our model on the price observed during this time interval. Then, using the estimated model, we investigate how well it predicts price from the last third of the auction, i.e. from the validation set. In other words, for auction i let

$$\mathcal{T}_i := \{(t_{is}, \text{Price}_{is}^{(1)}) | t_{is} < \frac{2}{3} * 7 \quad \text{days}\}$$

be the time/price pairs observed during the first two-thirds of the 7-day auction. This is the training data. Similarly, let

$$\mathcal{V}_i := \{(t_{is}, \text{Price}_{is}^{(2)}) | t_{is} \geq \frac{2}{3} * 7 \quad \text{days}\}$$

denote the validation data from the last auction-third. For comparison, we also investigate the performance of the penalized smoothing splines using the same approach. Since we cannot fit a penalized smoothing spline to auctions with less than 3 bids, we removed those auctions. This reduces the total set to 132 auctions.

We estimate both penalized splines and mixed model splines from the training data and compute the mean squared prediction error based on the validation data. Specifically, for the penalized splines we estimate the model

$$s(\text{Price}_{is}^{(1)}) = \alpha_0 + \boldsymbol{\phi}^T(t_{is}^{(1)})\boldsymbol{\alpha}_i$$

while in the case of the mixed models we estimate

$$s(\text{Price}_{is}^{(1)}) = \tilde{\alpha}_0 + \boldsymbol{\phi}^T(t_{is}^{(1)})\boldsymbol{\alpha} + b_{i0} + \boldsymbol{\phi}^T(t_{is}^{(1)})\tilde{\boldsymbol{\alpha}}b_i.$$

The mean squared prediction error is shown in Table 2. We can see that the penalized splines result in an MSE almost 60 times larger than that of the mixed model approach. This implies that taking a traditional smoothing approach can result in forecasts that are severely off.

	MSE
Penalized Spline	1,701,507
Mixed Model	28,352

Table 2: Mean squared prediction error for penalized spline and mixed model forecasting, respectively.

5 Conclusion

Functional data analysis often arrives with many data-related problems and challenges. One such challenge is sparse and unevenly distributed data. Traditional smoothing approaches often break down and/or produce conceptually not very meaningful results when the functional objects are sampled sparsely and unevenly. We propose a new approach which is based on the concept of mixed model methodology. In particular, we propose semiparametric mixed models together with boosting for parameter estimation. Our approach has several appeals: First, by borrowing information from similar functional objects, we can overcome challenging sparse data situations with as little as only one sample point per functional object. Moreover, our approach also allows to capture dependencies across functional objects. This is especially appealing in situations like ours where different processes (i.e. auctions) are hardly independent of one another. And lastly, by assuming a common underlying trend for all functional objects and by estimating this trend from all the data, our approach can induce shape restrictions on the functional objects without explicitly assuming any model-constraints. Our boosting approach allows for a convenient joint estimation of all model and smoothing parameters under one roof. The resulting model is parsimonious in that it adds only two additional parameters: the variance of the slope and the covariance between slope and intercept. It is very flexible, yet easy to interpret, which may make it an uncomplicated and pragmatic model for functional data.

On the substantive side, we contribute to the literature on online auctions by suggesting a new way of accounting for dependencies across different auctions. Much of the current online auction literature assumes independence which is typically not out of ignorance of the fact, but due the lack of appropriate statistical models. Online auction data feature complicated dependency structures: Auctions for the same (or similar) product may be correlated because they are competing for the same set of bidders. Moreover, repeat auctions by the same seller may be similar in terms of auction

design (e.g. auction length, opening bid, usage of a secret reserve price, number of pictures, quality of descriptions, etc.). This similarity in turn may lead to similar auction outcomes. And lastly, bidders have the freedom to participate in more than one auction. As a consequence, events in one auction (e.g. stark price increases) may cause bidders to updated their strategies in other auctions and thus the bids a bidder places in one auction are no longer independent from the bids s/he places in another auction. All of this means that online auction data can feature complicated dependencies. Our approach is one step into capturing some of these dependency structures.

A Algorithmic details

The algorithmic details of boosting are given below. For additional details see Tutz and Reithinger (2005).

BoostMixed

---

1. *Initialization*

   Compute starting values $\hat{\alpha}_0^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}$ and set $\hat{\eta_{\mathbf{i}}}^{(0)} = \mathbf{X}_i\boldsymbol{\delta}^{(0)}$, $\hat{\mathbf{Z}}_i^{(0)} = [1, \boldsymbol{\Phi}\hat{\boldsymbol{\alpha}}^{(0)}]$

2. *Iteration*

   For l=1,2,...

   (a) *Refitting of residuals*

   i. Computation of parameters

   One fits the model for residuals

   $$\mathbf{y}_i - \eta_{\mathbf{i}}^{(\mathbf{l-1})} \sim \mathbf{N}(\eta_{\mathbf{i}}, \mathbf{V}_{\mathbf{i}}^{(\mathbf{l-1})})$$

   with $\mathbf{V}_i^{(l-1)} = \mathbf{V}_i(\hat{\boldsymbol{\rho}}^{(l-1)}, (\hat{\sigma}_\epsilon^2)^{(l-1)}) = (\hat{\mathbf{Z}}_i^{(l-1)})^T\mathbf{Q}(\hat{\boldsymbol{\rho}}^{(l-1)})\hat{\mathbf{Z}}_i^{(l-1)} + (\hat{\sigma}_\epsilon^2)^{(l-1)}\mathbf{I}$ and $\eta_{\mathbf{i}} = \mathbf{X_i}\boldsymbol{\delta}$, yielding $\hat{\boldsymbol{\delta}}$.

   ii. Stopping step

   Stop if $BIC^{(l-1)}$ was smaller than $BIC^{(l)}$.

   iii. Update

Update for $i = 1, \ldots, n$ using $\hat{\boldsymbol{\delta}}^T = (\hat{\alpha}_0, \hat{\boldsymbol{\alpha}})$

$$\begin{aligned}
\boldsymbol{\eta}_i^{(l)} &= \boldsymbol{\eta}_i^{(l-1)} + \mathbf{X}_i \hat{\boldsymbol{\delta}}, \\
\hat{\boldsymbol{\alpha}}^{(l)} &= \hat{\boldsymbol{\alpha}}^{(l-1)} + \hat{\boldsymbol{\alpha}} \\
\hat{\alpha}_0^{(l)} &= \hat{\alpha}_0^{(l-1)} + \hat{\alpha}_0
\end{aligned}$$

and set

$$\mathbf{Z}_i^{(l)} = [1, \boldsymbol{\Phi}_i \boldsymbol{\alpha}^{(l)}].$$

(b) *Computation of Variance Components*

The computation is based on the penalized log-likelihood

$$\begin{aligned}
l_p(\boldsymbol{\theta}|\boldsymbol{\eta}^{(l)}; \delta^{(l)}) \quad &= -\tfrac{1}{2} \textstyle\sum_{i=1}^n \log(|\mathbf{V}_i^{(l)}|) + \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\eta}_i^{(l)})^T \mathbf{V}_i^{(l)} (\boldsymbol{\rho}, \sigma_\epsilon^2))^{-1} (\mathbf{y}_i - \boldsymbol{\eta}_i^{(l)}) \\
&\quad -\tfrac{1}{2} (\hat{\boldsymbol{\delta}}^{(l)})^T \mathbf{K} \hat{\boldsymbol{\delta}}^{(l)}.
\end{aligned}$$

Maximization yields $\boldsymbol{\rho}^{(l)}, (\sigma_\epsilon^2)^{(l)}$.

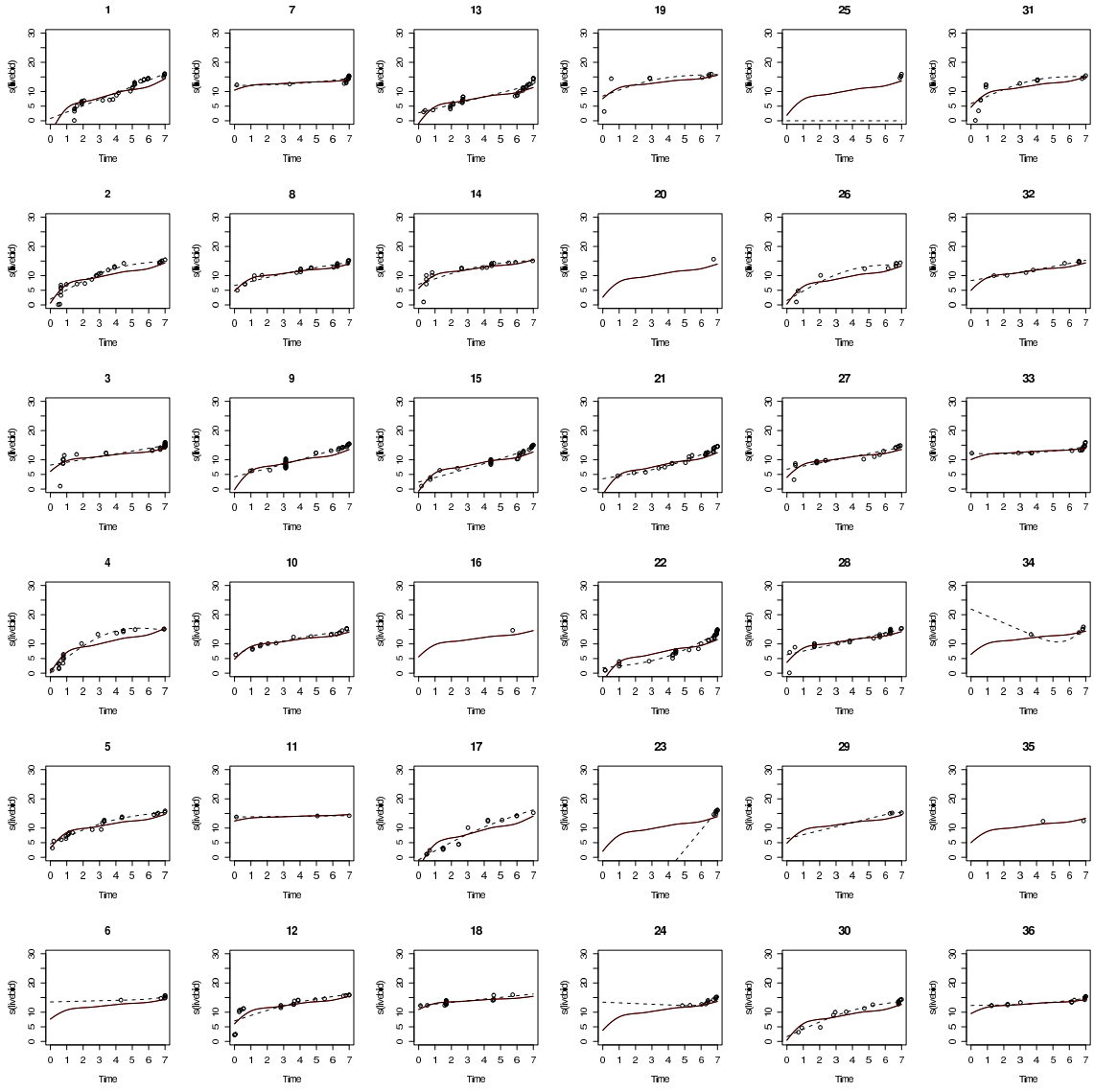


Figure 5: Smoothed Time: The first 36 auctions with their specific behavior regarding price and Time. Mixed model approach is shown by the solid lines, separately fitted penalized splines are the dotted lines.

References

Bapna, R., Goes, P., and Gupta, A. (2003). Analysis and design of business-to-consumer online auctions. *Management Science*, 49:85–101.

Bapna, R., Jank, W., and Shmueli, G. (2005). Price formation and its dynamics in online auctions. Technical report, Smith School of Business, University of Maryland, College Park.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11:1493–1517.

Brumback, B. A. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93:961–976.

Bühlmann, P. and Yu, B. (2003). Boosting with l2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied Statistics* (to appear).

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:337–407.

Ghani, R. (2005). Price prediction and insurance for online auctions. In the *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL, 2005.

Ghani, R. and Simmons, H. (2004). Predicting the end-price of online auctions. In the *Proceedings of the International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management*, Pisa, Italy, 2004.

Green, D. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–338.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

Hastie, T. and Tibshirani, R. (2000). Bayesian backfitting. *Statistical Science*, 15(4).

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, 9:226–252.

Jank, W. and Shmueli, G. (2005a). Modeling concurrency of events in online auctions via spatio-temporal semiparametric models. Technical report, Smith School of Business, University of Maryland, College Park.

Jank, W. and Shmueli, G. (2005b). Profiling price dynamics in online auctions using curve clustering. Technical report, Smith School of Business, University of Maryland.

Jank, W., Shmueli, G., and Wang, S. (2006). Dynamic, real-time forecasting of online auctions via functional models. In *The Proceedings of the Twelfth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD2006)*.

Kauffman, R. J. and Wood, C. A. (2003). Why does reserver price shilling occur in online auctions? In *Proceedings of the 2003 International Conference on Electronic Commerce.*

Krishna, V. (2002). *Auction Theory*. Academic Press, San Diego.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38:963–974.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society*, B61:381–400.

Lucking-Reiley, D., Bryan, D., Prasad, N., and Reeves, D. (1999). Pennies from ebay: the determinants of price in online auctions. Technical report, University of Arizona.

Marx, D. B. and Eilers, P. (1998). Direct generalized additive modelling with penalized likelihood. *Comp. Stat. & Data Analysis*, 28:193–209.

McCulloch, C. and Searle, S. (2000). *Generalized, Linear, and Mixed Models*. Wiley.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, linear and mixed models*. Wiley, New York.

Parise, H., Wand, M. P., Ruppert, D., and Ryan, L. (2001). Incorporation of historical controls using semiparametric mixed models. *Applied Statistics*, 50:31–42.

Ramsay, J. O. (1998). Estimating smooth monotone functions. 60(2):365–375.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*. Springer-Verlag, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag New York, 2nd edition.

Roth, A. E. and Ockenfels, A. (2002). Last-minute bidding and the rules for ending second-price auctions: Evidence from ebay and amazon auctions on the internet. *The American Economic Review*, 92(4):1093–1103.

Ruppert, D. and Carroll, R. J. (1999). Spatially-adaptive penalties for spline fitting. *Australian Journal of Statistics*, 42:205–223.

Ruppert, D. , Wand, M.P. and Carrol, R.J. *Semiparametric Regression*. Cambridge University Press.

Schimek, M. (2000). *Smoothing and Regression. Approaches, Computation and Application*. Wiley, New York.

Shmueli, G., Jank, W., Aris, A., Plaisant, C., and Shneiderman, B. (2006). Exploring auction databases through interactive visualization. *Decision Support Systems*. Forthcoming.

Shmueli, G., Russo, R. P., and Jank, W. (2004). Modeling bid arrivals in online auctions. Technical report, Working paper, Smith School of Business, University of Maryland.

Tutz, G. and Reithinger, F. (2005). Flexible semiparametric mixed models. SFB discussion paper 448. *SFB386*.

Verbeke, G. and Molenberghs, G. (2001). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. (1999). The anlysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics*, 48:269–311.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics*, 15:443–462.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, 18:223–249.

Wang, S., Jank, W., and Shmueli, G. (2006). Forecasting ebay's online auction prices using functional data analysis. *Journal of Business and Economic Statistics*. Forthcoming.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratric penalties. *Journal of the Royal Statistical Society B*, B(62):413–428.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additve models. *Journal of American Statistical Association*, 99:673–686.

Ruppert,D. ,Wand, M.P., Carroll R.J. Semi-parametric regression (2003) Cambridge University Press

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semi-parametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93:710–719.