Adaptive Nonparametric Regression for Marketplace Response Detection

Junqing Wu[*] Wendy Meiring[†] Yuedong Wang[‡]

Abstract
In today's technology world, data is increasingly being relied upon for decision-making. Nonparametric techniques have advantages in these scenarios because they don't enforce a lot of structure to the pattern in the data upfront. We have proposed a methodology that is highly adaptive in capturing sudden changes without overfitting. The adaptivity of the methodology comes from the use of multiple "libraries", such as spline and fourier bases, and the parsimony of model representation comes from ad-hoc estimation of model complexity via Monte Carlo simulations. Bootstrap confidence intervals can be obtained simultaneously as a by-product of the computation. It is shown that such a methodology has the potential to make a big impact in A/B testing and marketplace response detection, both of which are common statistical exercises in the technology industry. An R package has been made to implement the methodology.

Key Words: Basis selection, Generalized cross-validation, Generalized degrees of freedom, Nonparametric regression, Overcomplete basis, Spatial adaptivity, Smoothing spline ANOVA, Business Intelligence

1. Introduction

In the tech industry, especially for online businesses, A/B testing has become increasing common for comparing the outcomes of observation studies known as live experiments. Both the "A" and "B" in A/B testing respectively represent a combination of settings. A group of subjects are first divided into two groups supposedly matched in propensity of interest. Then these two groups undergo what is called an A/A test when both groups are subject to the "control" settings to make sure their respective outcomes in the metrics of interest are not significantly different. Afterwards one of the groups will be subject to the "treatment" settings and both groups will be observed for an equal period of time. The two groups are not always equal in size. So it usually is an unbalanced 2-factor design. Table 1 illustrates such a design. Factor A has two levels: control and treatment; Factor B can have a variable number of levels for different granularities in time, e.g. days or hours.

An ANOVA table can be generated and power of the test can be computed. The problem with this approach is that because of resource constraints, the experiments may not be run long enough to achieve the desirable statistical power and there is too much noise in the observations. To this end, we are proposing an adaptive non-parametric modeling procedure that will allow us to do a visual examination of the trend and variability over time in such data sets with high resolution.

[*] Marketplace Manager, Revenue at Microsoft Advertising, One Microsoft Way, Redmond, WA 98052, wjqpu@yahoo.com

[†] Associate Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106

[‡] Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106

Table 1: A/B testing: unbalanced 2-factor design

	B: Time j = 1, . . . , b		
A: Control i=0	. . .	$y_{(0,j,1)}, \ldots, y_{(0,j,n_{0j})}$	. . .
A: Treatment i=1	. . .	$y_{(1,j,1)}, \ldots, y_{(1,j,n_{1j})}$	. . .

2. Adaptive nonparametric methods

Consider a regression model

$$y_i = f(x_i) + \epsilon_i, \quad x_i \in \mathcal{X}, \, i = 1, \cdots, n, \tag{1}$$

where y_i's are the responses, x_i's are design points, f is a nonparametric function on an arbitrary domain $\mathcal{X}$ and ϵ_i's are iid random errors with mean zero. The goal is to model and estimate the mean function f.

Spline modeling is one popular approach to model the function f in (1). First consider the special case when $\mathcal{X} = [a, b]$. A spline is a piecewise function on intervals connected at fixed points called knots: $a < t_1 < \cdots < t_k < b$. There are two kinds of splines in general: regression splines and smoothing splines. Regression splines are usually considered a basis approach while the smoothing splines are considered a regularization approach. With a basis approach, one often starts with a set of linearly independent functions that span the function space. The number of basis functions is in general infinite. The task is to select a finite subset of basis functions that can approximate the function f well. With a regularization approach, one penalizes complexity and finds the optimal solution within a well-defined function space.

Both the basis and regularization approaches have a large literature and are commonly used in practice (Ruppert, Wand and Carroll 2003). Multivariate Adaptive Regression Splines, or MARS, developed by Friedman (1991), is an example of the basis approach. The smoothing spline analysis of variance, or SS-ANOVA (Wahba 1990), is an example of the regularization approach. Hybrid Adaptive Splines, or HAS (Luo and Wahba 1997), combines features from both regression spline and smoothing spline (hence the term "hybrid"). It is an example of both.

2.1 Limitation of current methods

Two key ingredients in non-parametric modeling are the choice of basis functions (or a model space) and a model selection procedure that selects basis functions and/or controls the balance between goodness-of-fit and model complexity. With respect to these two key ingredients, there are a few limitations with most of the existing methodologies.

All basis approaches mentioned earlier use one class of basis functions only. For example, MARS uses truncated polynomials, and HAS uses a basis generated by a reproducing kernel of the cubic splines for the univariate case. Even though a single class of basis functions, often of infinite dimension, may eventually be able to capture a complex signal in the data, the methods based on a single class of basis functions are limited in their adaptivity with finite samples. On the other hand, the choices for the class of basis functions are virtually unlimited.

As for model selection, for example, penalized least squares criterion relies on a single global smoothing parameter, λ, to control the trade-off between the goodness-of-fit and the smoothness of the estimated function over the entire domain $x \in [a, b]$. Thus an implicit assumption is that f is smooth with relatively homogeneous curvatures over the entire domain. If the true function is spatially inhomogeneous, then spline estimates tend to over-smooth in regions where f is rough and under-smooth in regions where f is smooth.

2.2 Generalized Degree of Freedom

The *Generalized Degrees of Freedom* (GDF) offers a way to adaptively estimate model complexity and is thoroughly investigated by Ye (1998). It is applicable to complex modeling procedures. The definition is based on the sum of the sensitivity of each fitted value to perturbation in the corresponding observed value. The concept is non-asymptotic in nature and offers a unified framework under which the complexity of any modeling procedure can be measured.

GDF depends on both the modeling procedure and the underlying true model. In a linear model, let

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' \tag{2}$$

be the projection matrix onto the space spanned by the columns of $\mathbf{X}$. Let $\hat{\boldsymbol{\mu}}(\mathbf{y}) = (\hat{\mu}_1, \ldots, \hat{\mu}_n) = \mathbf{Hy}$ be the fit. The degrees of freedom can be expressed as

$$p = tr(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i}. \tag{3}$$

The degrees of freedom are the sum of the sensitivities of the fitted values $\hat{\mu}_i$ with respect to the observed response values y_i.

For any modeling procedure $\mathcal{M}$, the GDF is defined by

$$D(\mathcal{M}) = \sum_{i=1}^{n} h_i^{\mathcal{M}}(\boldsymbol{\mu}), \tag{4}$$

where

$$h_i^{\mathcal{M}}(\boldsymbol{\mu}) = \frac{\partial E_\mu\{\hat{\mu}_i(\mathbf{y})\}}{\partial \mu_i} = \lim_{\delta\to 0} E_\mu \left\{ \frac{\hat{\mu}_i(\mathbf{y} + \delta \mathbf{e}_i) - \hat{\mu}_i(\mathbf{y})}{\delta} \right\} = \frac{1}{\sigma^2} cov(\hat{\mu}_i(\mathbf{y}), y_i), \tag{5}$$

where $\boldsymbol{\mu} = (Ey_1, \ldots, Ey_n)'$, $\mathbf{y} = (y_1, \ldots, y_n)'$, and $\mathbf{e}_i$ is the ith column of the $n \times n$ identity matrix.

2.3 The BSML procedure

As alluded to earlier, a truly flexible procedure should be able to choose from a variety of basis functions of different constructions to model complex features of a regression function. Let $\mathcal{L}_0$ and $\mathcal{L}_l = \{\psi_{l,1}, \ldots, \psi_{l,n_l}\}$ for $l = 1, \ldots, L$ be $L + 1$ libraries of basis functions. $\mathcal{L}_0$ is the null library, which could be empty or contains elements $\{\phi_1, \ldots, \phi_m\}$ that will be forced into the model. Each library contains basis functions with similar properties. Otherwise, there is no limitation on the number of libraries and the elements of each library. For example, there may be different families of basis functions for each $\mathcal{L}_l$: e.g. Fourier, wavelets or smoothing spline.

The idea of multiple libraries is inspired by the use of overcomplete bases in machine learning literature (Lewicki and Sejnowski 2000, Coifman and Wickerhauser 1992, Mallat and Zhang 1993, Chen, Donoho and Saunders 2001). With overcomplete bases, it is possible to represent a signal with better resolution and much smaller set of basis functions, leading to better interpretation of the model. But an overcomplete basis could also lead to overfitting if the corresponding rise in model complexity is not properly accounted for with the improvement in goodness-of-fit. That is where GDF comes to help.

BSML-C Procedure

1. *Initialization:* set $\mathcal{B}_m = \mathcal{L}_0$ and let M be the upper limit on the number of basis functions to be selected (including those in $\mathcal{L}_0$).

2. *Forward selection:* for $k = m+1, \ldots, M$, find ϕ_k from the remaining basis functions not yet selected in $\mathcal{O} = \cup_{l=1}^{L} \mathcal{L}_l$ that maximizes the reduction of residual sum of squares
$$\phi_k = \underset{\psi \,\in\, \mathcal{O} \,\cap\, \mathcal{B}_{k-1}^c}{\text{argmax}} \; \{RSS(\mathcal{B}_{k-1}) - RSS(\mathcal{B}_{k-1} \cup \{\psi\})\}$$
where $\mathcal{B}_{k-1}$ is the set of basis functions selected up to step k-1 and update $\mathcal{B}_k = \mathcal{B}_{k-1} \cup \{\phi_k\}$.

3. *Elimination:* choose $k^*, m \leq k^* \leq M$, as the minimizer of a model selection criterion $Cr(k)$ such as the GCV.

4. *Final model:* fit a standard or ridge regression model to the selected basis functions set $\mathcal{B}_{k^*}$.

BSML-S Procedure

1. *Initialization:* set $\mathcal{B}_m = \mathcal{L}_0$ and let M be the upper limit on the number of basis functions to be selected.

2. *Forward selection:* for $k = m + 1, \ldots, M$, select a candidate $\psi_{l,j_l}^{(k)}$ from the remaining basis functions not yet selected in library $\mathcal{L}_l, l = 1, \ldots, L$ that maximizes the reduction in the residual sum of squares
$$\psi_{l,j_l}^{(k)} = \underset{\psi \in \mathcal{L}_l \cap \mathcal{B}_{k-1}^c}{\text{argmax}} \; \{RSS(\mathcal{B}_{k-1}) - RSS(\mathcal{B}_{k-1} \cup \{\psi\})\}$$
where $\mathcal{B}_{k-1}$ is the set of basis functions selected up to step $k-1$. Then select $\phi_k \in \left\{\psi_{1,j_1}^{(k)}, \ldots, \psi_{L,j_L}^{(k)}\right\}$ that minimizes a model selection criterion $Cr_F(k)$ and update $\mathcal{B}_k = \mathcal{B}_{k-1} \cup \{\phi_k\}$.

3. *Elimination:* choose $k^*, m \leq k^* \leq M$ that minimizes a model selection criterion $Cr(k)$ such as the GCV.

4. *Final model:* fit a standard or ridge regression model to the the basis functions set $\mathcal{B}_{k^*}$.

More details regarding these two procedures are provided in Sklar, Wu, Meiring and Wang (2013) and Wu (2011).

2.4 Bootstrap confidence intervals

Confidence intervals can be obtained as a by-product of GDF computation. A bootstrap procedure for deriving confidence intervals generally follows the steps below. We first obtain a pilot fit $(\hat{\mu}_1^0, \hat{\mu}_2^0, \ldots, \hat{\mu}_n^0)$. With a Gaussian assumption for noise, we generate new residuals $\delta_1, \delta_2, \ldots, \delta_n$ from $N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2$ is the error variance estimated from the pilot fit or a difference-based estimator such as the Rice estimator. Alternatively, we can resample the residuals with replacement if the normal assumption is objectionable. Add them to the pilot fit and repeat T times so that we have T bootstrap samples

$$y_i^t = \hat{\mu_i^0} + \delta_i^t, \, i = 1, \ldots, n \quad t = 1, \ldots, T \tag{6}$$

where n is the sample size and T is the number of bootstrap samples. We use the BSML modeling procedure again to fit each bootstrap sample and obtain estimates $(\hat{\mu}_1^t, \hat{\mu}_2^t, \ldots, \hat{\mu}_n^t), \ t = 1, \ldots, T$. Then we can find the $(1-\alpha)\%$ (across the function) confidence interval for μ_i: $(\hat{\mu}_{\alpha/2,\, i}, \hat{\mu}_{(1-\alpha/2),\, i})$, where $\hat{\mu}_{\alpha/2,\, i}$ and $\hat{\mu}_{(1-\alpha/2),\, i}$ are the $\alpha/2$ and $(1-\alpha/2)$ percentiles of $(\hat{\mu}_i^1, \hat{\mu}_i^2, \ldots, \hat{\mu}_i^T)$.

Algorithm

1. Obtain a pilot fit from a “big” model: $(\hat{\mu}_1^0, \hat{\mu}_2^0, \ldots, \hat{\mu}_n^0)$ and an estimate of σ^2: $\hat{\sigma}^2$.

2. Generate bootstrap samples $y_i^t = \hat{\mu_i^0} + \delta_i^t, \, i = 1, \ldots, n, \, t = 1, \ldots, T$ where $\delta_1^t, \delta_2^t, \ldots, \delta_n^t$ are from $N(0, \hat{\sigma}^2)$ or a sample with replacement from the estimated residuals.

3. Apply the forward selection part of the BSML procedure to select up to M basis functions for each bootstrap sample and obtain all the fits: $(\hat{\mu}_{1,\, k}^t, \hat{\mu}_{2,\, k}^t, \ldots, \hat{\mu}_{n,\, k}^t)$, $k = m, \ldots, M$, $t = 1, \ldots, T$, where k indicates the number of basis functions in the model and m is the number of null bases.

4. Calculate $\text{MSE}(k, t) = \sum_{i=1}^n (\hat{\mu}_{i,\, k}^t - \hat{\mu}_i^0)^2/n, \ k = m, \ldots, M, \ t = 1, \ldots, T$ and find $k_t^* = \text{argmin}_{k=m}^M \text{MSE}(k, t), \ t = 1, \ldots, T$. This is to find the optimal fit for each bootstrap sample based on MSE. Since the true values are unknown, the values from the pilot fit are used as approximations.

5. Obtain the $(1-\alpha)\%$ confidence interval for $\mu_i, i = 1, \ldots, n$: $(\hat{\mu}_{\alpha/2,\, i}, \hat{\mu}_{(1-\alpha/2),\, i})$ from the range of estimates: $(\hat{\mu}_{i,\, k_1^*}^1, \hat{\mu}_{i,\, k_2^*}^2, \ldots, \hat{\mu}_{i,\, k_T^*}^T)$.

While bootstrap confidence intervals are not known for having the best properties in terms of coverage and width, it is found in Wu (2011) that, when used with the BSML procedure, bootstrap confidence intervals can have superior properties. A theoretical justification is being investigated.

2.5 The R package

The BSML procedure has been made into an R package (Wu, Sklar, Wang and Meiring 2012) available on CRAN with the following content:

- Main functions
  - bsml.R
  - bsmlc.R
  - bsmls.R
  - has.R
- Utility functions
  - control.bsml.R

- summary.bsml.R
- predict.bsml.R
- bas.check.R
- stdz.R

3. Applications

3.1 Marketing initiative effectiveness detection

In online advertising, when pushing a marketing initiative to advertisers, it is important to be able to detect whether an advertiser has taken action as suggested. One such example would be account managers reaching out to big advertisers whose accounts are underperforming (meaning getting no ad impressions) due to low bids. These advertisers are advised to raise the bids of their listings. Despite that advertisers are asked to call back within a week to confirm their activities (if action has been taken), the only way to know for sure is via the data logs. For a time period of 27 consecutive weeks, we calculate the proportion of listings whose bids have been raised for each account week over week: $\xi_{i,j,k}$ where $i = 0, 1$ (0 for control and 1 for treatment), $j = 1, \ldots, 26$ and $k = 1, \ldots, n_{ij}$ (n_{ij} are the number of advertisers in group i for week j). If we plot the mean of these proportions each week over time, for both control and treatment, we will get Fig. 1. The plot doesn't look very illuminating even though we can see in general the level of response for treatment is higher than that for control. If we are to model treatment and control separately: $\xi_{i,j,k} = f_i(j) + \epsilon_{ijk}$, where $i = 0, 1$, $j = 1, \ldots, 26$, $k = 1, \ldots, n_{ij}$ and ϵ_{jk} are iid random errors with mean zero, we can apply the BSML-C procedure to each. The libraries used were $\mathcal{L}_0 =$ constant and linear functions, $\mathcal{L}_1 =$ cubic spline representers, $\mathcal{L}_2 =$ truncated constants, $\mathcal{L}_3 =$ truncated linear polynomials and $\mathcal{L}_4 =$ truncated quadratic polynomials. The model selection criterion $Cr(k)$ used was GCV and 300 bootstrap samples were generated for computing the confidence bands. The outcome is a much clearer pattern for either control or treatment along with bootstrap confidence intervals showing the variability at each time point. In Fig. 2, it becomes clear that regardless of treatment (outreach), advertisers generally begin raising their bids at the beginning of the summer shopping season and lower their bids at the end of it. The start of the increase of activities was much more obvious for advertisers in the treatment group and they tend to continue to raise their bids throughout the shopping season. Similar observations were in fact made via other channels. Because of the treatment group was a much smaller group than control, the general lack of overlap between the confidence bands gives assurance that the difference in activities for the two groups indeed exists.

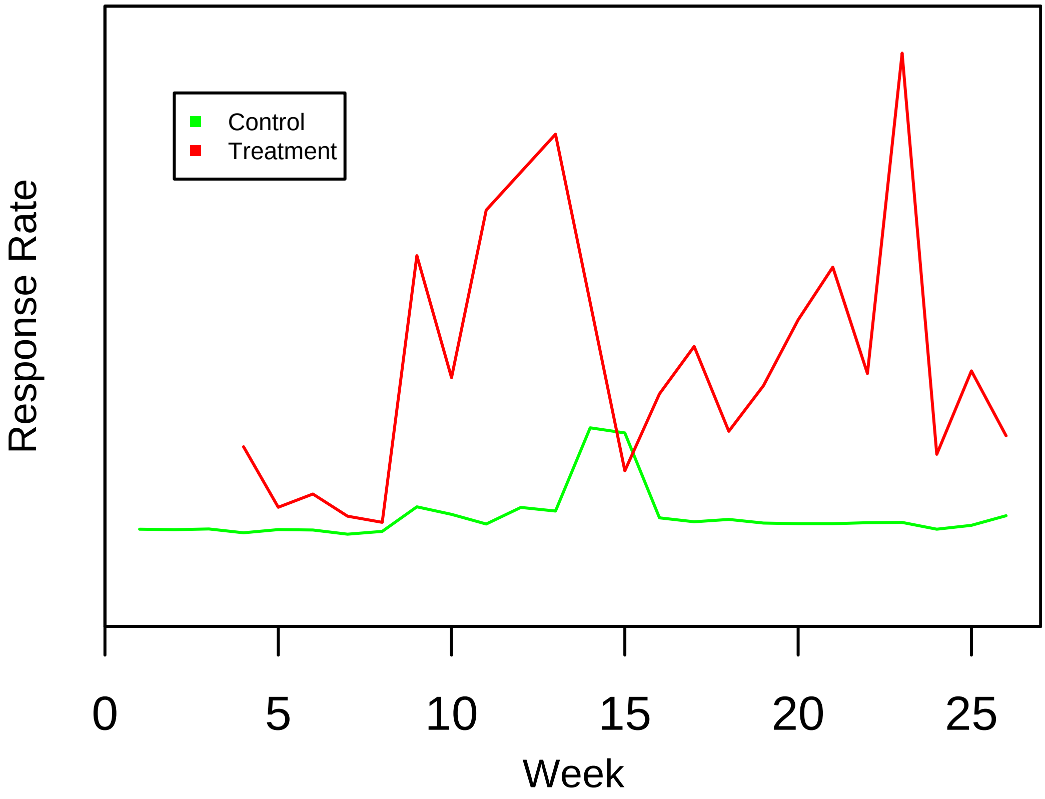

Figure 1: Average week-over-week market responses over time.

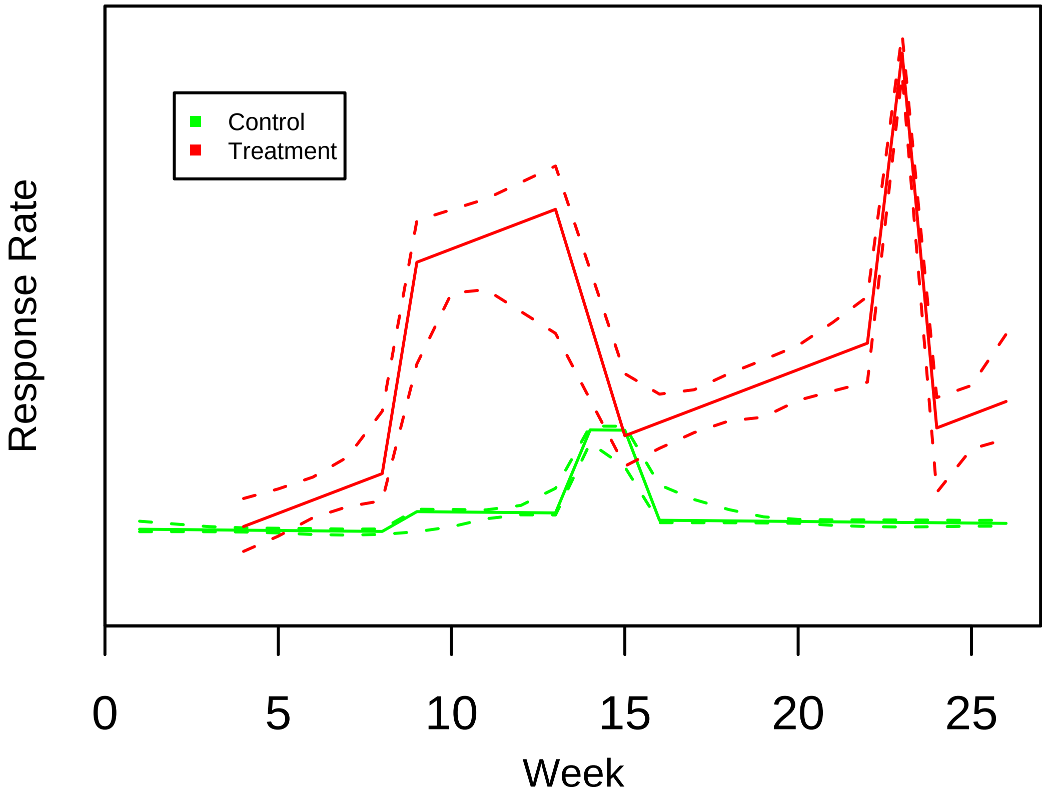

Figure 2: BSML fit to week-over-week market responses over time.

3.2 Positional factors estimation

The BSML procedure can be used to estimate the click-through rate (CTR) at each ad position all at once in online advertising. There are as many as 12 ad positions that can be filled on a search results page (though not all of them may be filled for every page) and they are clustered into two main blocks. The CTR values tend to be monotonically decreasing within each block from top to bottom and from the center block to the righthand block. When comparing the level of user engagement of one live experiment to that of another, we usually look at if the CTR values of one experiment are generally higher than those from another. In other words, we mostly care if one CTR "curve" is above another. The CTR values for each position also tend to be normally distributed. Therefore, the BSML procedure will be a great tool for this purpose, allowing detection of anomaly with its adaptivity while also providing bootstrap confidence intervals.

To show that there is an advantage to be had when computing bootstrap confidence intervals with BSML compared to computing them with other existing methods, as an example, in Fig. 3, we fit the CTR data using smoothing spline and also generated bootstrap confidence bands. There is sizable overlap between the confidence bands, making us wonder if the treatment curve is indeed above the control one. In Fig. 4, we are showing the BSML-C fit together with bootstrap confidence bands. Just as in the previous example, the libraries used were $\mathcal{L}_0 =$ constant and linear functions, $\mathcal{L}_1 =$ cubic spline representers, $\mathcal{L}_2 =$ truncated constants, $\mathcal{L}_3 =$ truncated linear polynomials and $\mathcal{L}_4 =$ truncated quadratic polynomials. The model selection criterion $Cr(k)$ used was GCV and 300 bootstrap samples were generated for computing the confidence bands. Because the confidence bands are much narrower, at the same time with as good or better coverage as shown in Wu (2011), we can clearly see that for positions 2, 3 and 4, all of which very important positions for revenue, the treatment curve is above the control curve. For the control curve in that figure, we can also see a reversal of monotonicity between positions 3 and 4, an important finding related to pricing that is not uncovered in the figure for smoothing spline. Such a finding is only possible because of the adaptivity and good balance between fit and model complexity from the BSML procedure.

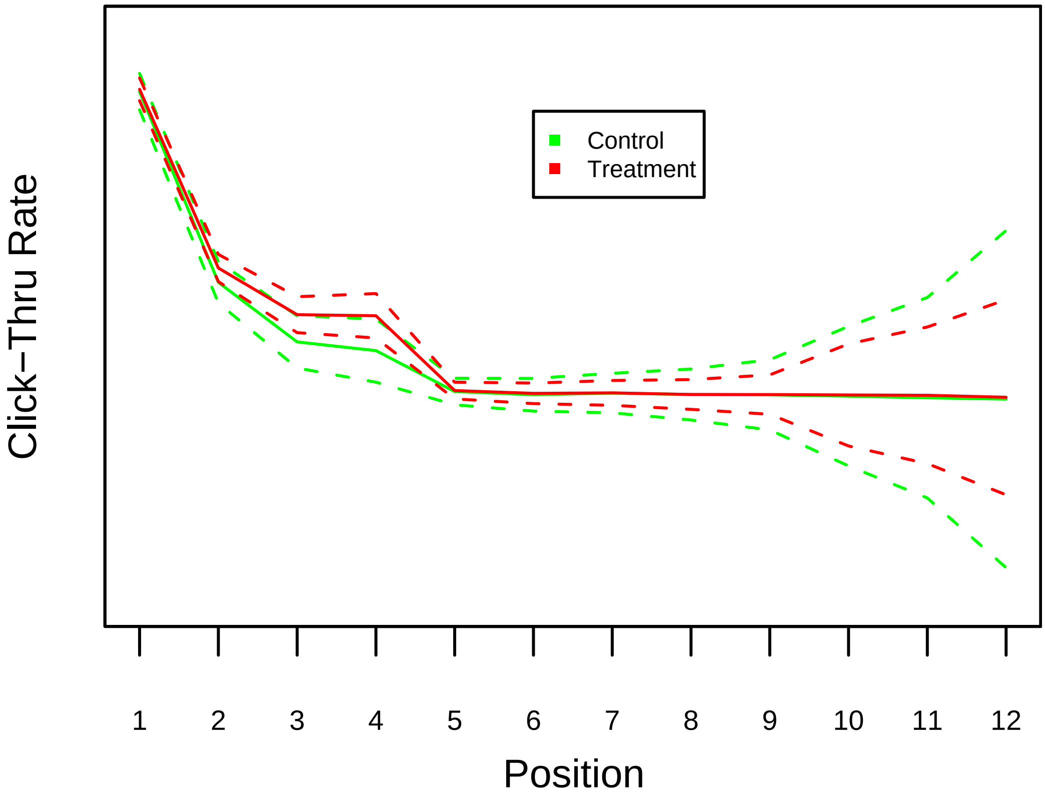

Figure 3: Smoothing spline fit to the CTR curves for both treatment and control.

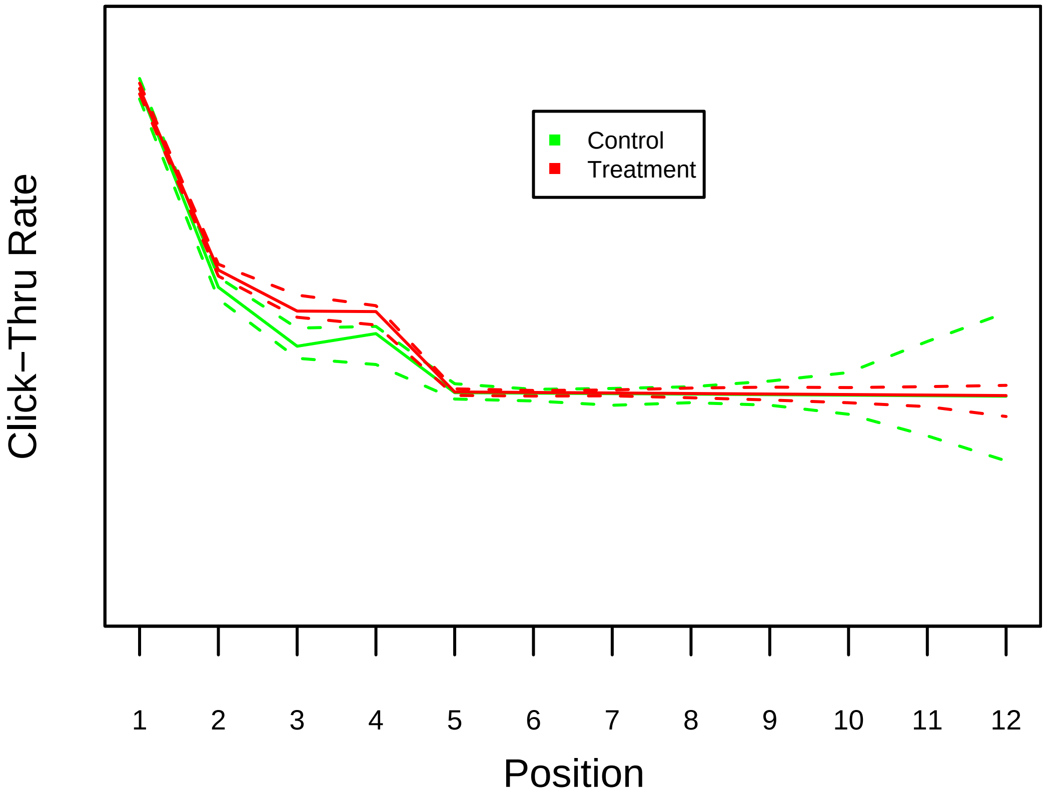

Figure 4: BSML-C fit to the CTR curves for both treatment and control.

4. Discussion

In this work we have illustrated the advantage of identifying the pattern in a live experiment using adaptive non-parametric procedures as well as comparing these patterns using graphical means. Any conclusions drawn from observing these graphical representations, however, will be rather subjective. Progress needs to be made for formally testing the difference between two curves. In addition, even though sample size generally is not an issue for the purpose of making the normal assumption, there are occasions either when the normal assumption doesn't hold or when we want to model the outcomes directly (e.g. 0 or 1). There is definitely a need to extend the BSML procedure to accommodate observations from the exponential family of distributions. Such an effort is underway.

References

Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001). Atomic decomposition by basis pursuit, *SIAM Review* **43**: 129–159.

Coifman, R. R. and Wickerhauser, M. V. (1992). Entropy-based algorithms for best-basis selection, *IEEE Trans. Inform. Theory* **38**: 713–718.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**: 1–67.

Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations, *Neural Computation* **12**: 337–365.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines, *Journal of the American Statistical Association* **92**: 107–116.

Mallat, S. and Zhang, Z. (1993). Matching pursuit in a time-frequency dictionary, *IEEE Trans. Signal Proc.* **41**: 3397–3415.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge, New York.

Sklar, J., Wu, J., Meiring, W. and Wang, Y. (2013). Non-parametric regression with basis selection from multiple libraries, *Technometrics*.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wu, J. (2011). Basis selection from multiple libraries, Ph.D. Thesis, University of California-Santa Barbara, Dept. of Statistics and Applied Probability.

Wu, J., Sklar, J. C., Wang, Y. and Meiring, W. (2012). *bsml: Basis Selection from Multiple Libraries*. R package version 1.5-1.

Ye, J. M. (1998). On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association* **93**: 120–131.