Clustering Curves in the Presence of Heteroscedastic Errors

Nicoleta Serban[1]
Industrial Systems and Engineering School
Georgia Institute of Technology
nserban@isye.gatech.edu

The clustering technique introduced in this paper is a means for discovering underlying patterns among a large number of curves. One novel characteristic compared to the current clustering methods is that we allow for heteroscedastic errors. Both the mean and the variance functions of each curve are assumed unknown and varying over time. The clustering method consists of a series of steps: transformation using an orthonormal basis of functions in L_2, dimension reduction through coefficient estimation in the transform domain, and clustering in the transform space. We show that in the transform space, the coefficient estimation procedure introduced in this paper is asymptotically optimal in the Pinsker's minimax sense over Sobolev ellipsoids. We illustrate our technique by clustering a large number of curves both within a synthetic example and within a specific application. In this application, we analyze the research and development expenditure over time of a subset of companies in the Compustat Global database. We show that our method compares favorably to two alternative approaches.

Key words and phrases: heteroscedastic regression, means model, modulation estimator, minimax optimal estimator, clustering multiple curves, the Compustat Global database.

1 Introduction

The primary objective in this paper is to cluster curves by shape. The clustering technique introduced here extends the procedure in Serban and Wasserman (2005)- CATS - to models with heteroscedastic errors. CATS provides a means for efficiently clustering curves by smoothing the curves, screening out flat curves, and clustering the non-constant smoothed curves. But CATS assumes that the variance of each

[1]The author is grateful to Larry Wasserman for his input in some of the proofs in this paper.

curve is constant over time. Constant variance is rather a restrictive assumption in applications where we observe a large number of curves or time series. Consequently, we assume that both the mean and the variance functions of each curve are unknown and non-constant over time.

The clustering method presented in this paper consists of a series of steps. First, we reduce the heteroscedastic nonparametric regression problem to a means estimation problem by transforming from the functional space to a transform space via an orthonormal basis of functions in L_2. Second, we estimate the coefficients or the means in the transform space using linear (modulation) estimators where the class of modulators satisfies a uniform entropy condition as defined in Beran and Dümbgen (1998). Finally, we cluster the estimated coefficients using the Euclidean distance to identify clusters of curves similar in shape. Euclidean clustering in the Fourier domain is equivalent to clustering using the correlation measure in function space. Pearson correlation is the most common measure for shape similarity.

Similar to the results in Beran and Dümbgen (1998), we show that for the heteroscedastic means model, the linear estimator that minimizes the Stein Unbiased Risk Estimator (SURE) achieves the asymptotic minimax bound. Since we use the Fourier basis to transform from the function space to the means model, optimal estimation in the transform space translates to optimal dimension reduction of the Fourier coefficients. Furthermore, the linear estimator that minimizes SURE is adaptive to unknown smoothness level in the sense that it approaches an "oracle" estimator assuming an asymptotic consistent estimator for the variance function in the function space. Adaptivity of the estimation procedure is essential here because we have to estimate the means for a large number of curves that have different noise levels, and different mean and variance functions.

Efromovich and Pinsker (1996) obtained similar results for minimax optimality and adaptive estimation under heteroscedastic errors where the nonparametric regression is transformed to a "white-noise" model. By comparison, we transform to a means model. We rather transform to a means model to use the means or the coefficients for clustering curves by shape.

For our technique, one model assumption is uncorrelated errors. For testing the assumption of uncorrelated errors, we extend the Durbin-Watson test for autocorrelation to heteroscedastic errors.

We investigate the performance of our clustering technique using a synthetic example. We compare our technique with two other methods: (1) Model-based method introduced in Yeung et al. (2001); and (2) The COSA method introduced in Friedman and Meulman (2004). The former method allows for different geometric cluster shapes (different covariance structures). The latter method uses a weighted distance, where the weights regulate the relative influence of the attributes describing the objects to be clustered. We also apply our clustering method to a real example where we study the temporal patterns of research and expenditure for companies listed in the Compustat Global database.

2 Model

We assume the nonparametric regression model:

$$Y_{ij} = s_i(t_{ij}) + \sigma_i(t_{ij})\epsilon_{ij} \text{ where } i = 1, \ldots, N,\ j = 1, \ldots, m_i \tag{1}$$

with N the number of curves and m_i the number of time points for i^{th} curve. Thus, Y_{ij} is the j^{th} observation on the i^{th} curve. Both functions $s_i(t)$ and $\sigma_i(t)$ are unknown. We want to estimate $s_i(t)$ while treating $\sigma_i(t)$ as a nuisance parameter estimated by an asymptotic consistent estimator $\hat{\sigma}_i(t)$ for all $i = 1, \ldots, N$.

We assume that the curves s_i belong to a Sobolev space $\mathcal{F} \equiv \mathcal{F}_\beta(c)$ of unknown order β and unknown radius c:

$$\left\{ s(x) = \sum_{j=1}^{\infty} \beta_j \psi_j(x) : \ \sum_{j=1}^{\infty} \beta_j^2 j^{2\beta} \leq c^2 \right\}$$

where $\psi_1, \psi_2, \ldots$ is an orthonormal basis for L_2.

We assume $\mathbb{E}(\epsilon_{ij}) = 0$ and the errors ϵ_{ij} are identically distributed and uncorrelated.

3 Transformation

We transform the regression nonparametric model to the means model using a basis of functions ϕ_i in L_2 as follows

$$Z_{ij} = \frac{1}{m_i} \sum_{k=1}^{m_i} Y_{ij}\phi_j(t_k) \text{ where } i = 1, \ldots, N,\ j = 1, \ldots, m_i.$$

For this transformation, the expectation of Z_{ij} can be approximated by

$$\mathbb{E}[Z_{ij}] = \frac{1}{m_i} \sum_{k=1}^{m_i} s_i(t_k)\phi_j(t_k) \approx \int f_i(t)\phi_j(t)dt := \theta_{ij}$$

and the variance of Z_{ij} can be approximated by

$$\mathbb{V}[Z_{ij}] = \frac{1}{m_i^2} \sum_{k=1}^{m_i} \sigma_i^2(t_k)\phi_j^2(t_k) \approx \frac{1}{m_i} \int \sigma_i^2(t)\phi_j^2(t)\, dt := \frac{\gamma_{ij}^2}{m_i}. \tag{2}$$

Moreover, if we sum up the variances in (2), we can further write

$$\sum_{j=1}^{m_i} \gamma_{ij}^2 = \sum_{j=1}^{m_i} \int \sigma_i^2(t)\phi_j^2(t)\,dt = \int \sigma_i^2(t) \sum_{j=1}^{m_i} \phi_j^2(t)\,dt = \int \sigma_i^2(t)\Phi_{m_i}^2(t)\,dt \tag{3}$$

where $\Phi_m^2(x) := \sum_{i=1}^m \phi_i^2$. We define the means model as

$$Z_{ij} = \theta_{ij} + \frac{\gamma_{ij}}{\sqrt{m_i}} E_{ij}, \text{ for } j = 1, \ldots, m_i. \tag{4}$$

where θ_{ij} and γ_{ij} are both unknown. Here $i = 1, \ldots, N$ is the curve index.

Under the independence assumption of ϵ_{ij}, the errors E_{ij} are random variables identically distributed with median 0. Often we assume that $E_i \sim N(0, 1)$.

In our clustering method, we will use the cosine basis of functions when s_i are aperiodic and the cosine-sine basis when s_i are periodic for $i = 1, \ldots, N$.

4 Modulation

For ease of presentation, we will drop the curve index while keeping in mind that the procedure we are about to describe applies to all curves. Consider the general means problem:

$$Z_i = \theta_i + \frac{\gamma_i}{\sqrt{n}} E_i, \ i = 1, \ldots, n.$$

Given an estimator $\hat{\theta}$ of θ, let the loss function be

$$L(\hat{\theta}, \theta) = \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2 = ||\hat{\theta} - \theta||^2. \tag{5}$$

The corresponding risk function is given by

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta(L(\hat{\theta}, \theta)). \tag{6}$$

We consider the class of linear estimators $\hat{\theta} = fZ$, where $f \in \mathcal{F} \subset [0, 1]^n$. These estimators are called modulation estimators and f is called modulator in Beran and Dümbgen (1998). More specifically, in our clustering technique, the class of modulators is

$$f \in \mathcal{M}_n = \{(1, 0 \ldots, 0), (1, 1, 0 \ldots, 0), (1, 1 \ldots, 1)\}, \tag{7}$$

and we find the linear estimator that minimizes the risk function over $\mathcal{M}_n$. Using the class of modulators $\mathcal{M}_n$, the estimation in the transform space translates to dimension reduction in the sense that only the first few estimated coefficients or means are further used to explain the shape information in a curve. But the results in this paper hold for any class for modulators $\mathcal{F}$ for which

$$J(\mathcal{F}) = \int_0^1 \sqrt{\log(N(u,\mathcal{F}))} du < \infty,$$

where $N(u,\mathcal{F})$ is the uniform covering number of $\mathcal{F}$.

For linear estimators, the risk function is given by

$$R(f,\theta,\sigma) = \sum_{i=1}^{n} \left(\frac{\gamma_i^2}{n} f_i^2 + \theta_i^2 (1-f_i)^2 \right). \tag{8}$$

We cannot use (8) to optimize for f because the means θ_i are unknown. To find the optimal estimator, we use an estimated version of the risk. To this end, we use the Stein's unbiased risk estimator (SURE):

$$\hat{R}(f) = \sum_{i=1}^{n} \left(\frac{\hat{\gamma}_i^2}{n} f_i^2 + (Z_i^2 - \frac{\hat{\gamma}_i^2}{n})(1-f_i)^2 \right). \tag{9}$$

The following result is used to show the optimality and adaptivity of the linear estimator than minimizes $\hat{R}(f)$ for the heteroscedastic model.

Theorem 1 *Let $\mathcal{F}$ be any closed subset of $[0,1]^n$ containing 0. Define*

$$\tilde{f} = \arg\min_{f\in\mathcal{F}} R(fZ,\theta,\sigma) \text{ and } \hat{f} = \arg\min_{f\in\mathcal{F}} \hat{R}(f).$$

Assume $\theta \in \Theta(m,c)$, the Sobolev ball of radius c. Then the following inequalities hold:

$$\mathbb{E}|L(\hat{f}Z,\theta) - R(\tilde{f},\theta,\sigma)| \le B(\gamma,\hat{\gamma},\theta,\mathcal{F}), \tag{10}$$

$$\mathbb{E}|\hat{R}(\hat{f}) - R(\tilde{f},\theta,\sigma)| \le B(\gamma,\hat{\gamma},\theta,\mathcal{F}), \tag{11}$$

$$\mathbb{E}|R(\hat{f},\theta,\sigma) - R(\tilde{f},\theta,\sigma)| \le B(\gamma,\hat{\gamma},\theta,\mathcal{F}), \tag{12}$$

where the common upper bound is given by

$$B(\gamma,\hat{\gamma},\theta,\mathcal{F}) = CJ(\mathcal{F}) \left(\sqrt{\sum_{i=1}^{n} \frac{\gamma_i^4}{n^2}} + \sqrt{\sum_{i=1}^{n} \frac{\gamma_i^2}{n}\theta_i^2} \right) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|\hat{\gamma}_i^2 - \gamma_i^2|. \tag{13}$$

The proof, which follows closely the proof of Theorem 2.1 in Beran and Dümbgen (1998), can be found in the Appendix.

The following result shows that the modulation estimator $\hat{\theta} = \hat{f}Z$ for the heteroscedastic model is optimal in the minimax sense over the Sobolev ellipsoids.

Corollary 1 *The linear estimator $\hat{f}Z$ achieves the asymptotic minimax bound provided the following conditions hold:*

1. *In the function space, the variance function $\sigma^2 \in L_2[0,1]$,*

2. *The modulator set $\mathcal{F}$ satisfies $J(\mathcal{F}) = o(n^{1/2})$.*

3. *The variance estimator under the means model is $\hat{\gamma}_i^2 = \int_0^1 \hat{\sigma}^2(t)\phi_i^2(t)dt$, where $\hat{\sigma}^2$ is a consistent estimator of the variance function $\sigma^2(t)$.*

See the Appendix for the proof of Corollary 1.

The following result shows that the linear estimators for the heteroscedastic model is adaptive for means in the Sobolev balls.

Theorem 2 *Under the same notations and assumptions in Theorem 1 and Corollary 1, we have*

$$\mathbb{E}\left[\sum_{i=1}^{n}(\hat{f}_i Z_i - \tilde{f}_i Z_i)^2\right] \leq B(\gamma, \hat{\gamma}, \theta, \mathcal{F}) \tag{14}$$

and

$$\mathbb{E}\left[\sum_{i=1}^{n}(\frac{\gamma_i^2}{n} + \theta_i^2)(\hat{f}_i - \tilde{f}_i)^2\right] \leq B(\gamma, \hat{\gamma}, \theta, \mathcal{F}). \tag{15}$$

In other words, $\hat{f}Z$ approaches $\tilde{f}Z$, the "oracle" modulation estimator for the set of modulators $\mathcal{F}$.

The proof of Theorem 2.2 follows closely the proof of Theorem 2 in Beran and Dümbgen (1998). See the Appendix for the proof of Theorem 2.

In our clustering technique, using the Fourier transforms, we further simply the problem as follows: Let $\theta = fZ$, where $f \in \mathcal{M}_n$ is defined in (7). Therefore, we only consider estimators with modulators of the form $f = (1, 1, \ldots, 1, 0, \ldots, 0)$. Or equivalently,

$$\hat{\theta}_i = \begin{cases} Z_i, & i \leq J. \\ 0 & \text{else} \end{cases}$$

Here J is referred to as the smoothing parameter.

For adaptive and optimal estimation, we estimate the smoothing parameter as below

$$\hat{J} = \text{argmin}_{J=2,\ldots,n} R(J),$$

where

$$R(J) = \left(\sum_{i=1}^{J}\frac{\hat{\gamma}_i^2}{n} + \sum_{i=J+1}^{n}\left(Z_i^2 - \frac{\hat{\gamma}_i^2}{n}\right)_+\right).$$

Rather than simply finding the minimum risk $R(J)$ over J, we estimate the smoothing parameter to be the shift point in the risk function. To illustrate, suppose we can write the risk function as

$$R(J) = \begin{cases} \alpha + \gamma J, & J \leq J_0, \\ (\alpha - \delta J_0) + (\gamma + \delta) J, & J \geq J_0. \end{cases}$$

We can estimate the change point in the risk function using the Bacon and Watts (1971) method, Farley and Hinch(1971) method or simply estimating J_0 by maximizing the likelihood function.

For the model in (4), we estimate the smoothing parameter $\hat{J}_i$ for each curve i. If $\hat{\theta}_i = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{im_i})$ are the estimated coefficients of curve i, then the estimated curve is:

$$\hat{s}_i(t) = \sum_{j=1}^{\hat{J}_i} \hat{\theta}_{ij} \phi_j(t).$$

The smoothing parameter will be shape-dependent. That is, $\hat{J}_i$ is the number of coefficients that explain most of the shape information in curve i. So we expect that curves similar in shape will have similar smoothing parameters.

5 Variance Estimation

For the minimax result to hold, we need a consistent estimator for the variance function $\sigma^2(t)$ where the model in the function space is

$$Y_i = s(t_i) + \sigma(t_i)\epsilon_i, \; i = 1, \ldots, n. \tag{16}$$

We use the estimator in Fan and Yao (1998), where local linear regression with bandwidth h_2 is applied to the squared residuals after estimating the mean function as in Fan (1993) using a bandwidth h_1. The variance function estimator is adaptive to the unknown regression function s and is efficient under the smoothness condition that the bandwidth h_2 converges to 0 no more slowly than h_1. We denote $\hat{\sigma}^2(t)$ and $\hat{s}(t)$ the local linear estimators of the variance function $\sigma^2(t)$, and, respectively, of the mean function $s(t)$ in Fan and Yao (1998).

Under mild regularity conditions, from Theorem 1 in Fan (1993) and Theorem 1 in Fan and Yao (1998), $s(t)$ is an asymptotic consistent estimator of $s(t)$ and $\hat{\sigma}^2(t)$ is an asymptotic consistent and efficient estimator of $\sigma^2(t)$ for optimal bandwidths h_1 and h_2 ($h_1 = O(n^{-1/5})$ and $h_2 = O(n^{-1/5})$). Optimal bandwidths can be selected as in Fan and Gijbels (1995) or Ruppert, Sheather and Wand (1995).

There is an extensive body of literature on the estimation of variance functions in nonparametric regresion. See Müller and Stadtmüller (1993), Müller and Zhao (1995), Härdle and Tsybakov (1997), Efromovich (1999) and the references therein.

6 Test for Autocorrelation of Heteroscedastic Errors

Since we assumed a model where the error term is uncorrelated, but the variance is non-constant over time, we shall check this assumption by testing for autocorrelation in the presence of heteroscedasticity.

For a general model in the function space in (16), we estimate $s(t)$ and $\sigma^2(t)$ as in Section 5. We define the standardized residual as

$$\hat{r}(t_j) = \frac{Y_j - \hat{s}(t_j)}{\hat{\sigma}(t_j)}.$$

The asymptotic distribution of $\hat{r}(t)$ does not depend on t. Therefore, we can employ standard techniques for testing the presence of lag-1 autocorrelation. A test statistic similar to the test in Durbin and Watson (1950) for lag-1 autocorrelation is

$$T(r) = \frac{\sum_{j=2}^{N}(r(t_j) - r(t_{j-1}))^2}{\sum_{j=1}^{N} r(t_j)^2}.$$

A statistic for a lag-m autocorrelation test is introduced in Wang (1998).

Rather than obtaining an approximate (asymptotic) distribution for $T(r)$ under the null hypothesis of uncorrelated errors, we estimate the p-values using nonparametric bootstrap. Since under the null hypothesis the errors ϵ_j are uncorrelated identical distributed, we can estimate the null distribution of $T(r)$ by sampling without replacement from $r(t_1), \ldots, r(t_N)$ B times and estimate the p-value as:

$$p = \frac{1}{B}\sum_{b=1}^{B} I(T(r) \geq T(r_b^*)).$$

To illustrate the coverage and the power of the modified Durbin-Watson test for autocorrelation in the presence of heteroscedastic errors, we simulate p-values under the null hypothesis:

$$Y_j = s(x_j) + .3\epsilon_j,\ \epsilon_j = \sigma(x_j)e_j,\ j = 1, \ldots, 200,$$

and under the alternative hypothesis:

$$Y_j = s(x_j) + .3\epsilon_j,\ \epsilon_j = .3\epsilon_{j-1} + \sigma(x_j)e_j,\ j = 1, \ldots, 200$$

where $s(x) = 2(x^2\sin(x) + (1 - x^2)\cos(x))$, $\sigma^2(x) = s\sin^2(x)$ and $e_j \sim N(0, 1/2)$. Figure 1 shows 200 estimated p-values under the null and alternative hypotheses. For this simulation, both the coverage and the power are high.

7 Clustering

Our primary objective in this paper is to cluster curves by shape. A common measure for shape similarity is Pearson correlation. But clustering using the correlation measure in the function space is equivalent to Euclidean clustering in the transform domain. Let $f = \sum_j a_j\phi_j$ and $g = \sum_j b_j\phi_j$ decompositions of curves f and g using an orthonormal basis. From $a = (a_1, a_2 \ldots)$ define a new vector $\tilde{a} = (\tilde{a}_2, \tilde{a}_3, \ldots)$ obtained by discarding a_1 and normalizing:

$$\tilde{a}_j = \frac{a_j}{\sqrt{\sum_{j=2}^{m} a_j^2}}, \quad j \geq 2. \tag{17}$$

Define $\tilde{b} = (\tilde{b}_1, \tilde{b}_2, \ldots)$ similarly. Then,

$$\rho(f, g) = 1 - \frac{\|\ \tilde{a} - \tilde{b}\ \|^2}{2}. \tag{18}$$

In our examples, we assigned the cluster membership using the hierarchical clustering algorithm. Any conventional clustering algorithm can be used here.

8 Synthetic Example

We generate synthetic data from the following regression model,

$$Y_{ij} = f_i(t_{ij}) + \sigma(t_{ij})\epsilon_{ij}, \tag{19}$$

where $t_{ij} = j/m$, $j = 1, \ldots, 100$, and $i = 1, \ldots, 3000$.

Figure 2 displays six different patterns for the mean functions f_i for $i = 1, \ldots, 3000$. The synthetic data consist of 500 curves for each of the six curve shapes. The error term ϵ_{ij} is assumed to be iid $N(0, 1)$. In equation (19), the variance also varies over time. The variance function for each curve is randomly selected out of eight variance patterns. See Figure 3 for the eight variance functions used in this synthetic example.

We estimate the smoothing parameter J_i for each synthetic curve $\{Y_{ij},\ j = 1, \ldots, m\}$. Almost all synthetic curves that are similar in shape to patterns F_2 or F_5 in Figure 2 have an estimated smoothing parameter of $\hat{J} = 9$, where synthetic curves simulated from patterns F_3 or F_6 in Figure 2 have an estimated smoothing parameter of $\hat{J} = 8$. Synthetic curves simulated from patterns F_1 or F_4 have either $\hat{J} = 8$ or $\hat{J} = 9$. Therefore, curves with similar patterns also have similar estimated smoothing parameters.

For each of the six patterns, we randomly selected a synthetic curve and plotted its estimated pattern and its true pattern in Figure 4. The estimated underlying patterns follow closely the true ones.

We compare different clustering techniques using a misclustering error rate defined as follows. Let $T_{n,K}$ and $\hat{T}_{n,K}$ denote the true clustering map, and, respectively, the estimated clustering map. We define the true clustering map as

$$T_{n,K}(f,g) = \begin{cases} 1 & \text{if } f \text{ and } g \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

We define the estimated clustering map $\hat{T}_{n,K}(f,g)$ similarly. The two clustering maps depend on the number of clusters.

The *misclustering rate* for K clusters is

$$\eta(K) = \frac{1}{\binom{N}{2}} \sum_{r<s} I\Big(T_{n,K}(f_r, f_s) \neq \hat{T}_{n,K}(f_r, f_s) \Big). \tag{21}$$

This clustering error rate can also be expressed as $\eta = 1 - \mathcal{R}(T, \hat{T})$ where $\mathcal{R}$ is the Rand index (Rand, 1971).

Now we compare our method with two clustering alternative methods:

1. The model-based clustering method introduced by Yeung et al. (2001) (*mclust*) applied to the raw synthetic data in the functional space.

2. The COSA clustering method introduced by Friedman and Meulman (2004), where the attributes are the m Fourier coefficients. COSA provides another way to select the coefficients (attributes) that best describe the underlying pattern in each cluster.

For this synthetic example, our clustering technique outperforms the other two methods. The misclustering rate $\eta(K)$ for $K = 6$ clusters is about .1 when applying model based clustering and about .06 when applying COSA to the synthetic example. The clustering method introduce in this paper identifies the true cluster memebership with zero error.

9 Application: Research and Development Expenditure

The motivating application for our clustering analysis is identifying patterns in the research and development expenditure for companies listed in the Compustat Global database. This database comprises financial and market data for about 20,500 US companies. The available variables are all those included in publicly disclosed financial statements, and their respective notes and restated values, when available.

We are particularly interested in the dynamics of the research and development expenditure (R&D) presented in millions of dollars. There are only about 440 companies that have quarterly inputs for R&D starting with year 1990 or earlier. Therefore, there is a very small percentage of companies in this large database that have more than 60 data inputs for R&D expenditure. We will analyze the pattern of the R&D expenditure for these 440 companies. Most of these companies are from Manufacturing industry sector.

First, we do not expect highly correlated time series since the data are observed quarterly. To validate our assumption we test for auto-correlation using the test introduced in Section 6. The estimated p-values are in Figure 5. Correcting for multiplicity using the False Discovery Rate, we reject the null hypothesis of uncorrelated observations for only two companies.

Second, we expect non-constant variability over time due to the large number of factors associated with R&D expenditure. The variability of these factors changes over time.

So we are within the model assumptions in Section 2. For most of the curves in our application, the estimated smoothing parameter is less than nine. So there is considerable dimension reduction. Nine randomly chosen observed curves and their estimates are shown in Figure 6.

We further apply the clustering algorithm to the smoothed Fourier (cosine) coefficients. We estimate about five clusters using the method introduced in Tibshirani, Walther, Hastie (2000). The mean curves in each cluster are in Figure 7. Most of the curves fall in the first cluster (315 out of 440). Hence the good news is that most of the company have an increasing expenditure on research and development. However, there are several companies with bell shaped R&D expenditure patterns such as the ones in clusters two, three and four. Among these companies are: LDM technologies, which was taken off the market in 2005; Koppers Inc, a large producer of aluminium, steel, and railroads, that may have budget problems due to the high maintenance expenditure; Cingular Wireless LLC, which increased its expenditure on research and development in 2004, the year in which it merged with AT&T wireless; and Regal Cinemas Inc, which appears to have invested in improving and opening new centers in the early 1990's.

This investigation of the patterns discovered in R&D expenditure is being incorporated in a larger study in which we are interested in the effect of R&D expenditure and other variables on sales and profits of the companies in the Compustat Global database.

10 Discussion

In this article, we present a procedure for clustering a large number of curves in the presence of heteroscedastic errors. The main contribution of the current method is that we allow for both the mean and the variance functions of each curve to vary over time. Two main advantages of using our procedure are that we employ an optimal dimension reduction via transformation from the functional space into the Fourier domain, and adaptivity to the unknown smoothness of the curves. It is important to reduce the dimensionality since we cluster a large number of curves and, therefore, we need an inexpensive computational method. For example, in our synthetic example we reduced from 100 dimensions in the function space to only about 9 dimensions in the Fourier domain. It is important to employ an estimation method that adapts to the unknown smoothness of the curves, since the curves will have different patterns in the mean and variance functions, and different noise levels.

In time series analysis, it is often the case that the noise component is correlated over time. We do not account for correlation in our method. Adaptive and optimal estimation under correlated errors remains a challenge especially under fixed design points (see Efromovich (1999)).

Nonetheless, there is a broad spectrum of applications that fall within our statistical framework. One example is introduced in this paper. In our motivating example, we are interested in identifying underlying patterns in research and development expenditure for companies in the Compustat Global database. The information about these patterns is further investigated in a larger study that will be discussed in a future research paper.

Appendix

Proof of theorem 1: Define $W_1(i) = E_i^2 - \gamma_i^2/n$, $W_2(i) = \theta_i E_i$ and $V_i = (\hat{\gamma}_i^2 - \gamma_i^2)/n$.

Following the notation above, we write:

$$L(fY, \theta) - R(f, \theta, \sigma) = \textstyle\sum_{i=1}^n \left(f_i^2 W_1(i) + 2(f_i^2 - f_i)W_2(i)\right)$$

and

$$\hat{R}(f) - R(f, \theta, \sigma) = \textstyle\sum_{i=1}^n \left((f_i^2 - 2f_i + 1)(W_1(i) + 2W_2(i)) + (2f_i - 1)V_i\right).$$

Define $\mathcal{G} = \{fg; f, g \in \mathcal{F}\}$. According to the equalities above we have

$$\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f, \theta, \sigma)| \leq 4 \sup_{g\in\mathcal{G}} \left|\sum_{i=1}^n g_i W_1(i)\right| + 8 \sup_{g\in\mathcal{G}} \left|\sum_{i=1}^n g_i W_2(i)\right| + \left|\sum_{i=1}^n V_i\right|.$$

The results in Theorem 1 follow from the next lemma.

Lemma 1 *Under the previous assumptions and notations, the following inequalities hold:*

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n} g_i W_1(i)\right| \le CJ(\mathcal{F})\sqrt{\sum_{i=1}^{n}\mathbb{E}(E_i^4)} \tag{22}$$

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n} g_i W_2(i)\right| \le CJ(\mathcal{F})\sqrt{\sum_{i=1}^{n}(\mathbb{E}(E_i^2)\theta_i^2)} \tag{23}$$

Proof of Lemma 1: We begin with the proof of Equation (22). Let

$$\psi_1(i)(g) = E_i^2 g_i,$$

where $S(g) = \sum_{i=1}^{n}\psi_1(i)(g)$. By a result in Beran & Dümbgen (1998), we have

$$\mathbb{E}||S-\mathbb{E}S||_{\mathcal{G}} \le C\int_0^{\hat{D}}\sqrt{\log N(u,\mathcal{G},\hat{\rho})}du \tag{24}$$

where $\hat{D} = \sup_{g\in\mathcal{G}}\hat{\rho}(g,0)$, assuming that $0\in\mathcal{F}$. In addition, we also have

$$\hat{\rho}(g,h)^2 = \sum_{i=1}^{n} E_i^4(g_i-h_i)^2 = \sum_{i=1}^{n} E_i^4\sum_{i=1}^{n}\frac{E_i^4}{\sum_{i=1}^{n}E_i^4}(g_i-h_i)^2 =$$

$$\sum_{i=1}^{n} E_i^4 d_{\hat{Q}}(g,h)^2 \le \sum_{i=1}^{n} E_i^4.$$

In the last equation, $\hat{Q}$ is a discrete distribution that puts weights $p_i = \frac{E_i^4}{\sum_{i=1}^{n}E_i^4}$ on time points $t_i\in[0,1]$. Therefore,

$$\hat{D} = \sup_{g\in\mathcal{G}}\hat{\rho}(g,0) \le \left(\sum_{i=1}^{n}E_i^4\right)^{1/2}.$$

Finally, let us rewrite

$$\sum_{i=1}^{n} g_i W_1(i) = \sum_{i=1}^{n} g_i E_i^2 - \sum_{i=1}^{n} g_i\mathbb{E}[E_i^2] =$$

$$\sum_{i=1}^{n}\psi_1(i)(g) - \mathbb{E}\left[\sum_{i=1}^{n}\psi_1(i)(g)\right] = S(g)-\mathbb{E}[S(g)].$$

By the inequality in (24), we therefore have

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n} g_i W_1(i)\right| \le C\left(\sum_{i=1}^{n}\mathbb{E}[E_i^4]\right)^{1/2}\int_0^1\sqrt{\log N(u,\mathcal{G})}du.$$

The proof of Equation (23) follows similar steps. All it takes is to change the first two lines to

$$\psi_2(i)(g) = \theta_i E_i g_i$$

and

$$\hat{\rho}^2(g,h) \le d_Q^2(g,h) \sum_{t\in T} \theta^2(t) E^2(t).$$

With this we end the proof of Lemma 1.

Lemma 2 *Under the means model, suppose that the variance estimator is given by $\hat{\gamma}_i^2 = \int_0^1 \hat{\sigma}^2(t)\phi_i^2(t)dt$, where $\hat{\sigma}^2(t)$ is a consistent estimator of $\sigma^2(t)$. Furthermore, we assume that the modulator set $\mathcal{F}$ satisfies $J(\mathcal{F}) = o(n^{1/2})$, and the variance function satisfies $\int_0^1 \sigma^2(t)\, dt < \infty$. Then the bound*

$$B(\gamma, \hat{\gamma}, \theta, \mathcal{F}) = CJ(\mathcal{F}) \left(\sqrt{\sum_{i=1}^{n} \frac{\gamma_i^4}{n^2}} + \sqrt{\sum_{i=1}^{n} \frac{\gamma_i^2}{n}\theta_i^2} \right) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}|\hat{\gamma}_i^2 - \gamma_i^2|$$

goes to zero as $n \to \infty$.

Proof of Lemma 2: We first find an upper bound for the first term of $B(\gamma, \hat{\gamma}, \theta, \mathcal{F})$:

$$CJ(\mathcal{F}) \left(\sqrt{\sum_{i=1}^{n} \frac{\gamma_i^4}{n^2}} + \sqrt{\sum_{i=1}^{n} \frac{\gamma_i^2}{n}\theta_i^2} \right).$$

Let $\theta \in \Theta(m, c)$ (the Sobolev ellipsoid of radius c) for $m > 1/2$ and $c > 0$. The sum in (3) imply

$$CJ(\mathcal{F}) \left(\sqrt{\sum_{i=1}^{n} \frac{\gamma_i^4}{n^2}} + \sqrt{\sum_{i=1}^{n} \frac{\gamma_i^2}{n}\theta_i^2} \right) \le CJ(\mathcal{F}) \left(\frac{1}{n}\sqrt{\int \sigma^4(t)\Phi_n^2(t)dt} + \frac{c^2}{\sqrt{n}}\sqrt{\int \sigma^2(t)\Phi_n^2(t)dt} \right) \tag{25}$$

as provided below. The first part in the inequality in (25) is derived from the Jensen inequality. First we use the Jensen inequality

$$\gamma_i^4 = \left(\int_0^1 \sigma^2(t)\phi_i^2(t)dt \right)^2 = \left(\mathbb{E}_{\phi_i^2}[\sigma^2(X)] \right)^2 \le \mathbb{E}_{\phi_i^2}[\sigma^4(X)] = \int_0^1 \sigma^4(t)\phi_i^2(t)dt$$

and then sum up

$$\sum_{i=1}^{n} \gamma_i^4 \le \int_0^1 \sigma^4(t)\Phi_n^2(t)dt. \tag{26}$$

The second part in the inequality in (25) follows from

$$\sum_{i=1}^{n}\gamma_i^2\theta_i^2=\int_0^1\sigma^2(t)\sum_{i=1}^{n}(\phi_i^2(t)\theta_i^2)dt=(\sum_{i=1}^{n}\theta_i^2)\int_0^1\sigma^2(t)\frac{\sum_{i=1}^{n}(\phi_i^2(t)\theta_i^2)}{\sum_{i=1}^{n}\theta_i^2}dt \tag{27}$$

and

$$c^2\int_0^1\sigma^2(t)\frac{\sum_{i=1}^{n}(\phi_i^2(t)\theta_i^2)}{\sum_{i=1}^{n}\theta_i^2}dt\leq c^2\int_0^1\sigma^2(t)\Phi_n^2(t)dt.$$

We now find the rate of convergence of the first part of the bound $B(\gamma,\hat{\gamma},\theta,\mathcal{F})$ as $n\to\infty$ using the inequality in (25). According to the inequality in (26), we have

$$\frac{1}{n^2}\sum_{i=1}^{n}\gamma_i^4\leq\frac{1}{n}\int\sigma^4(t)\frac{\Phi_n^2(t)}{n}\,dt\leq\frac{1}{n}\int_1^0\sigma^4(t)\,dt.$$

Hence the rate of convergence is

$$\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\gamma_i^4}=O(n^{-1/2})\text{ when }\int_0^1\sigma^4(t)\,dt<\infty. \tag{28}$$

Next, we find the asymptotic rate of

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\gamma_i^2\theta_i^2}.$$

The assumption $\theta\in\Theta(m,c)$ implies that

$$\frac{1}{n}\sum_{i=1}^{n}\gamma_i^2\theta_i^2\leq\frac{1}{n}\sum_{i=1}^{n}\frac{\gamma_i^2}{i^{2m}}\sum_{i=1}^{n}i^{2m}\theta_i^2\leq\frac{c^2}{n}\int_0^1\sigma^2(t)\sum_{i=1}^{n}\frac{\phi_i^2(t)}{i^{2m}}dt.$$

Defining

$$\Psi_n^2(t)=\sum_{i=1}^{n}\frac{\phi_i^2(t)}{i^{2m}}.$$

we have the limit as $n\longrightarrow\infty$

$$\int_0^1\Psi_n^2(t)dt=\sum_{i=1}^{n}\frac{1}{i^{2m}}\to\zeta(2m),$$

where $\zeta(2m)$ is the Riemann zeta function. When $n>1$ and $m>1/2$, we have

$$\zeta(n)=\frac{1}{\Gamma(n)}\int_0^\infty\frac{u^{x-1}}{e^u-1}du<\infty.$$

Consequently, the rate of convergence is

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\gamma_i^2\theta_i^2} = O(n^{-1/2}) \text{ when } \int_0^1 \sigma^2(t)\,dt < \infty. \tag{29}$$

Finally, we show that the second part in the bound $B(\gamma, \hat{\gamma}, \theta, \mathcal{F})$ is zero in limit as $n \to \infty$. Since $\tilde{\gamma}_i^2 = \frac{1}{n}\sum_{j=1}^{n}\sigma^2(t_j)$, we can write

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\hat{\gamma}_i^2 - \gamma_i^2| \le \frac{1}{n}\sum_{i=1}^{n}|\tilde{\gamma}_i^2 - \gamma_i^2| + \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}|\hat{\sigma}^2(t_j) - \sigma^2(t_j)|\frac{\Phi_n^2(t_j)}{n}.$$

The first term goes to 0 if $\sigma^2(t)$ is integrable since

$$\frac{1}{n}\sum_{j=1}^{n}\sigma^2(t_j)\phi_i^2(t_j) \longrightarrow \int_0^1 \sigma^2(t)\phi^2(t)\,dt$$

by the definition of the Riemann integral.

For the second term, we have

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{\gamma}_i^2 - \tilde{\gamma}_i^2| \le \frac{1}{n}\sum_{j=1}^{n}|\hat{\sigma}^2(t_j) - \sigma^2(t)|\frac{\Phi_n^2(t_j)}{n}$$

where, if we assume $|\phi_i(t)| \le 1$, we get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|\hat{\gamma}_i^2 - \tilde{\gamma}_i^2|] \le \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[|\hat{\sigma}^2(t_j) - \sigma^2(t_j)|] \approx \int_0^1 \mathbb{E}[|\hat{\sigma}^2(t) - \sigma^2(t)|]\,dt,$$

which converges to zero if $\hat{\sigma}^2(t)$ is a consistent estimator for $\sigma^2(t)$.

This solves the asymptotic of the second part of the bound $B(\gamma, \hat{\gamma}, \theta, \mathcal{F})$:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\hat{\gamma}_i^2 - \gamma_i^2| \longrightarrow_{n\to\infty} 0. \tag{30}$$

The rate of convergence of this second part of the bound depends on the rate of convergence of the consistent estimator $\hat{\sigma}^2(t)$ for $\sigma^2(t)$.

Proof of corollary 1: We first introduce the following notations:

$$\delta^2 = \inf_{\hat{\theta}} \sup_{\theta\in\Theta(\alpha,c)} R(\hat{\theta}, \theta, \gamma),$$

$$\nu^2 = \sup_{\theta\in\Theta(\alpha,c)} \inf_{f\in\mathcal{F}} R(fZ, \theta, \gamma)$$

$$\delta_L^2 = \inf_{f\in\mathcal{F}} \sup_{\theta\in\Theta(\alpha,c)} R(fZ,\theta,\gamma).$$

For the minimax risks above, we can show that $\delta^2 \leq \nu^2 \leq \delta_L^2$. According to Pinsker(1980), the limit

$$\frac{\delta_L^2}{\delta^2} \longrightarrow 1 \text{ holds as} \frac{n\nu^2}{\sup_i \gamma_i} \to \infty.$$

As stated in Lemma 2, we have

$$\mathbb{E}|R(\hat{f},\theta,\gamma) - R(\tilde{f},\theta,\gamma)| \leq CJ(\mathcal{F})O(n^{-1/2}) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\hat{\gamma}_i^2 - \gamma_i^2|.$$

So under the limit assumption, we have

$$\sup_{\theta\in\Theta(\alpha,c)} R(\hat{f}Z,\theta,\gamma) \leq \sup_{\theta\in\Theta(\alpha,c)} R(\tilde{f}Z,\theta,\gamma) + CJ(\mathcal{F})o(n^{-1/2})+$$

$$\sup_{\hat{\theta}\in\Theta(\alpha,c)} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|\hat{\gamma}_i^2 - \gamma_i^2| = \nu^2 + o(1)$$

Combining the previous derivation with the assumptions in this Corollary, we have

$$\frac{\sup_{\theta\in\Theta(\alpha,c)} R(\hat{f}Z,\theta,\gamma)}{\delta^2} \longrightarrow 1 \text{ when } \frac{n\sup_{\theta\in\Theta(m,s)}\inf_{\hat{\theta}=fZ} R(\hat{\theta},\theta,\gamma)}{\sup_j \gamma_j^2} \to \infty.$$

This shows that the estimator $\hat{f}Z$ is asymptotically minimax over the Sobolev ellipsoids.

Proof of Theorem 2: We first prove the inequality (14). Define

$$w_1(i) = Y_i^2, \quad w_2(i) = \gamma_i^2/n + \theta_i^2$$

and

$$\hat{g}_i = \frac{Z_i^2 - \hat{\gamma}_i^2/n}{Z_i^2} \quad \tilde{g}_i = \frac{\theta_i^2}{\gamma_i^2/n + \theta_i^2}.$$

Show first that:

$$f_1 = \hat{f} = \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} w_1(i)(f_i - \hat{g}_i)^2 \tag{31}$$

$$f_2 = \tilde{f} = \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} w_2(i)(f_i - \tilde{g}_i)^2 \tag{32}$$

where $\hat{f}$ defines:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f) = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \left[\frac{((1-f_i)Z_i^2 - \gamma_i^2/n)^2}{Z_i^2} + \gamma_i^2/n - \gamma_i^4/n^2 \right]$$

and $\tilde{f}$ defines:

$$\tilde{f} = \arg\min_{f \in \mathcal{F}} R(f, \theta, \sigma) = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \left[\frac{((1-f_i)\theta_i^2 - f_i\gamma_i^2/n)^2}{\theta_i^2 + \gamma_i^2/n} + \theta_i^2 - \frac{\theta_i^4}{\theta_i^2 + \gamma_i^2/n} \right].$$

Consider the quadratic function:

$$T_{21}(x) = \sum_{i=1}^{n} w_2(i) \left[(1-x)\tilde{f}_i + x\hat{f}_i - \tilde{g}_i \right]^2$$

Because we assume $\mathcal{F}$ convex, so $(1-x)\tilde{f} + x\hat{f} \in \mathcal{F}$, and using (32) we can further find that:

$$T_{21}(x) \geq \sum_{i=1}^{n} (\tilde{f}_i - \tilde{g}_i)^2 = T_{21}(0). \tag{33}$$

But $T_{21}(x)$ is quadratic, so $T'_{21}(0) \geq 0$ since $T_{21}(x) \geq T_{21}(0)$. We can write this equivalently as

$$2\sum_{i=1}^{n} w_2(i)(\tilde{f}_i - \tilde{g}_i)(\hat{f}_i - \tilde{f}_i) \geq 0 \tag{34}$$

and

$$T'_{21}(1) \geq 2\sum_{i=1}^{n} w_2(i)(\tilde{f}_i - \hat{f}_i)^2. \tag{35}$$

Similarly, when considering

$$T_{12}(x) = \sum_{i=1}^{n} w_1(i) \left[(1-x)\hat{f}_i + x\tilde{f}_i - \hat{g}_i \right]^2$$

obtain similar inequalities such as

$$2\sum_{i=1}^{n} w_1(i)(\hat{f}_i - \hat{g}_i)(\tilde{f}_i - \hat{f}_i) \geq 0. \tag{36}$$

Rewrite $T'_{21}(x)$ and use inequality (36)

$$T'_{21}(x)|_{x=1} = -T'_{21}(1-x)|_{x=0} = -2\sum_{i=1}^{n} w_2(t)(\tilde{f}_i - \hat{f}_i)(\hat{f}_i - \tilde{g}_i) \leq$$

$$2\sum_{i=1}^{n} w_1(i)(\hat{f}_i - \hat{g}_i)(\tilde{f}_i - \hat{f}_i) - 2\sum_{i=1}^{n} w_2(t)(\tilde{f}_i - \hat{f}_i)(\hat{f}_i - \tilde{g}_i) =$$

$$2\sum_{t\in T}\left[w_1(i)(\hat{f}_i - \hat{g}_i) - w_2(i)(\hat{f}_i - \tilde{g}_i)\right](\tilde{f}_i - \hat{f}_i). \quad (37)$$

Now from (35) and (37) we obtain the inequality in (38) which is one step to the proof of Theorem 2:

$$\sum_{i=1}^{n} w_2(i)(\hat{f}_i - \tilde{f}_i)^2 \leq \sum_{i=1}^{n}\left[w_1(i)(\hat{f}_i - \hat{g}_i) - w_2(i)(\hat{f}_i - \tilde{g}_i)\right](\tilde{f}_i - \hat{f}_i). \quad (38)$$

Rewrite

$$w_2(i)(\tilde{f}_i - \tilde{g}_i) - w_1(i)(\tilde{f}_i - \hat{g}_i) = \sum_{i=1}^{n}\left[\tilde{f}_i(\gamma_i^2/n + \theta_i^2 - Z_i^2) - \hat{\gamma}_i^2 - \theta_i^2 + Z_i^2\right] =$$

Replace $Z_i^2 = (E_i + \theta_i)^2$

$$\sum_{i=1}^{n}\left[\tilde{f}_i(\gamma_i^2/n - E_i^2 - 2E_i\theta_i) - (\hat{\gamma}_i^2/n - \gamma_i^2/n) - (\gamma_i^2/n - E_i^2) + 2E_i\theta_i\right] =$$

$$\sum_{i=1}^{n}\left[(1 - \tilde{f}_i)(W_1(i) + 2W_2(i)) - V_i\right].$$

The inequality (38) becomes:

$$\sum_{i=1}^{n} Z_i^2(\hat{f}_i - \tilde{f}_i)^2 \leq \sum_{i=1}^{n}\left[(1 - \tilde{f}_i)(W_1(i) + 2W_2(i)) - V_i\right](\hat{f}_i - \tilde{f}_i) \leq$$

$$2\sup_{f\in\mathcal{F}}|\sum_{i=1}^{n}\left[(1 - \tilde{f}_i)(W_1(i) + 2W_2(i))\right] f_i| + \sum_{i=1}^{n}|V_i| \leq$$

$$4\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n} g_i W_1(i)\right| + 8\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n} g_i W_2(i)\right| + \sum_{i=1}^{n}|V_i|$$

By Lemma 1 we obtain (32).

Similarly we obtain the inequality:

$$\sum_{i=1}^{n}(\frac{\gamma_i^2}{n} + \theta_i^2)(\hat{f}_i - \tilde{f}_i)^2 \leq \sum_{i=1}^{n}\left[(\hat{f} - 1)(W_1(i) + 2W_2(i)) + V_i\right](\tilde{f}_i - \hat{f}_i) \leq$$

$$4\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}(g_i W_1(i)\right| + 8\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}(g_i W_2(i)\right| + \sum_{i=1}^{n}|V_i|.$$

By Lemma 1 we obtain (31).

References

[1] Bacon, D., Watts, D. (1971), "Estimating the transition between two intersecting straight lines", *Biometrika*, 58, 525-534.

[2] Barndorff-Nielsen, O.E., Cox, D.R. (1989), *Asymptotic Techniques for Use in Statistics*, Chapman and Hall, London.

[3] Beran, R. (2000), "REACT Scatterplot Smoothers: Superefficiency through basis economy", *JASA*, 95, 449, pp 155-171.

[4] Beran, R., Dümbgen, L. (1998), "Modulation of estimators and confidence sets", *Annals of Statistics*, 26, 5, pp 1826-1856.

[5] Durbin, J., Watson, G.S. (1971), "Testing for serial correlation in least squares regression", *Biometrika*, 58, 1 - 19.

[6] Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer, NY.

[7] Efromovich, S., Pinsker, M. (1996), "Sharp-Optimal and Adaptive estimation for heteroscedastic nonparametric regression", *Statstica Sinica*, 6, 925-942.

[8] Fan, J. (1993), "Local Linear Regression Smoothers and their Minimax efficienty", *Ann. of Statistics*, 21, 196-216..

[9] Fan, J., Gijbels, I. (1995), "Data-driven Bandwidth Selection in Local Plynomial Fitting: Variable Bandwidth and Spatial Adaptation", *J. R. Statistical Society B*, 57, 371-394.

[10] Fan, J., Yao, Q. (1998), "Efficient estimation of conditional variance functions in stochastic regression", *Biometrika*, 85, 3, pp. 645-660.

[11] Farley, J., Hinch, M. (1970), "A test for a shifting slope coefficient in a linear model", *J. of the American Statistical Association*, 65, 1320-1329.

[12] Friedman, J., Meulman, J.J. (2004), "Clustering Objects on Subsets of Attributes", *Journal of the Royal Statistical Society, B*, 66, 815-845.

[13] Härdle, W., Tsybakov, A. (1997), "Local polynomial estimators of the volatility function in nonparametric autoregression", *J. of Econometrics*, 81, pp 223-242.

[14] Müller, H.G., Stadtmüller, U. (1993), "On variance function estimation with quadratic forms",*J. of Statistical Planning and Inferences*, 35, 213-231.

[15] Müller, H.G., Zhao, P.L. (1995), "On a semiparametric variance function model and a test for heteroscedasticity", *Ann. of Statistics*, 23, 946-967.

[16] Pisier, G. (1983), "Some applications of the metric entropy condition to harmonic analysis", *Banach spaces, Harmonic Analysis and Probability Theory.* Lecture Notes in Math., 995, 123-154, Springer, NY.

[17] Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods".*Journal of the American Statistical Association* , 66, pp. 846-850.

[18] Ruppert, D., Sheather, S.J., Wand, M.P. (1995), "An effective bandwidth selector for local least squares regression", *J. of American Statistical Association*, 90,1257-1270.

[19] Tibshirani, R., Walther, G., Hastie, T. (Dec 2000), "Estimating the number of clusters in a dataset via the Gap statistic", *Journal of the Royal Statistical Society, B*,2000.

[20] Yeung, K.Y., Murua, A., Raftery, A., Ruzzo, W.L. (2001). "Model-Based Clustering and Data Transformations for Gene Expression Data", *Bioinformatics*, 17, 977-987.

[21] Wang, Y.J. (1998), "A test of Autocorrelation in the Presence of Heteroskedasticity of Unknown Form", *Econometric Theory*, 14, 87-122.

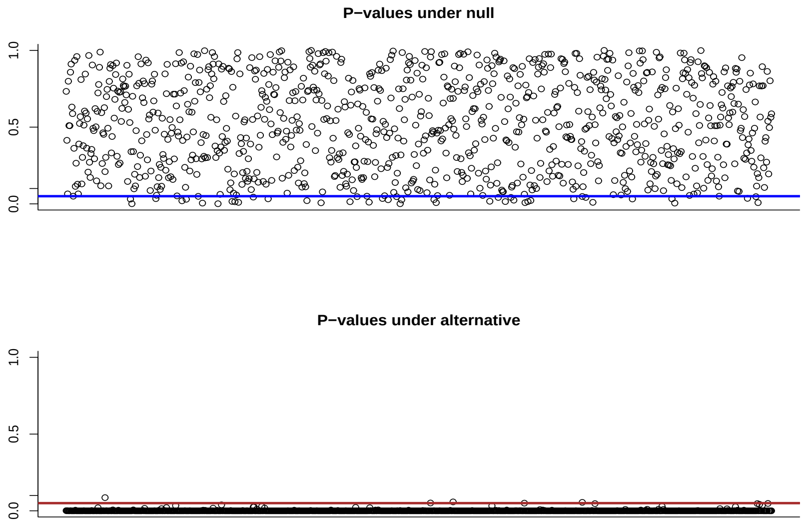

Figure 1: P-values under the null of uncorrelated heteroscedastic errors (upper plot) and p-values under the alternative of correlated heteroscedastic errors (bottom plot).

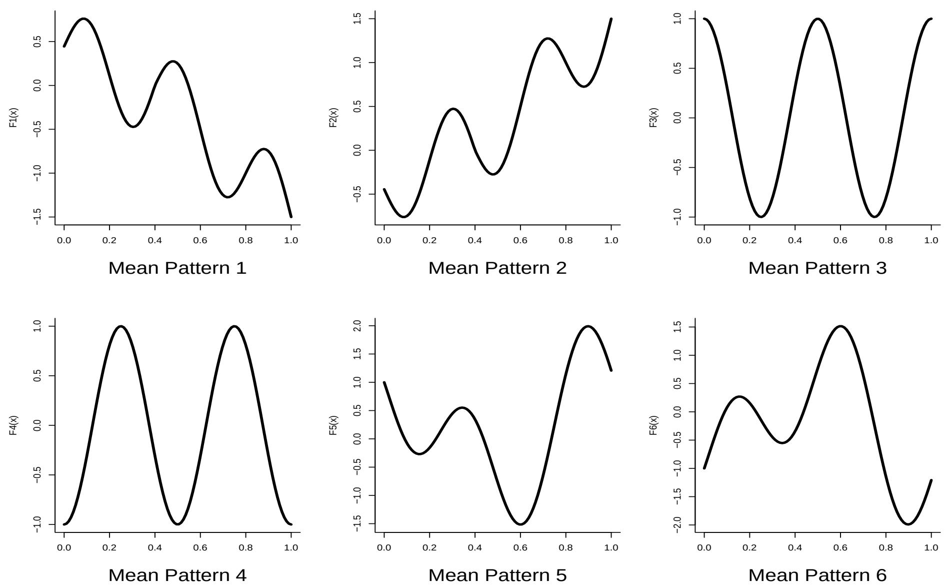

Figure 2: Six mean patterns for our synthetic example.

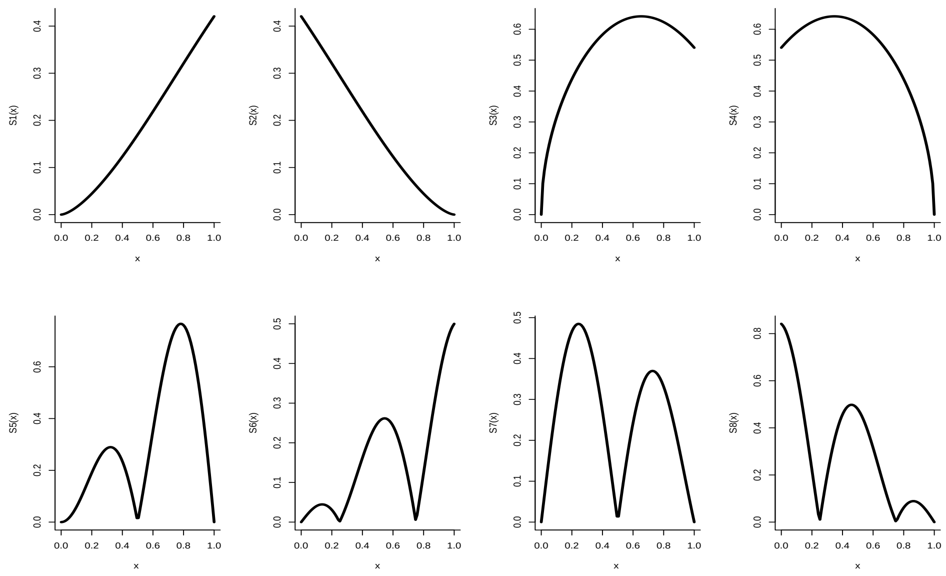

Figure 3: Eight variance patterns for our synthetic example.

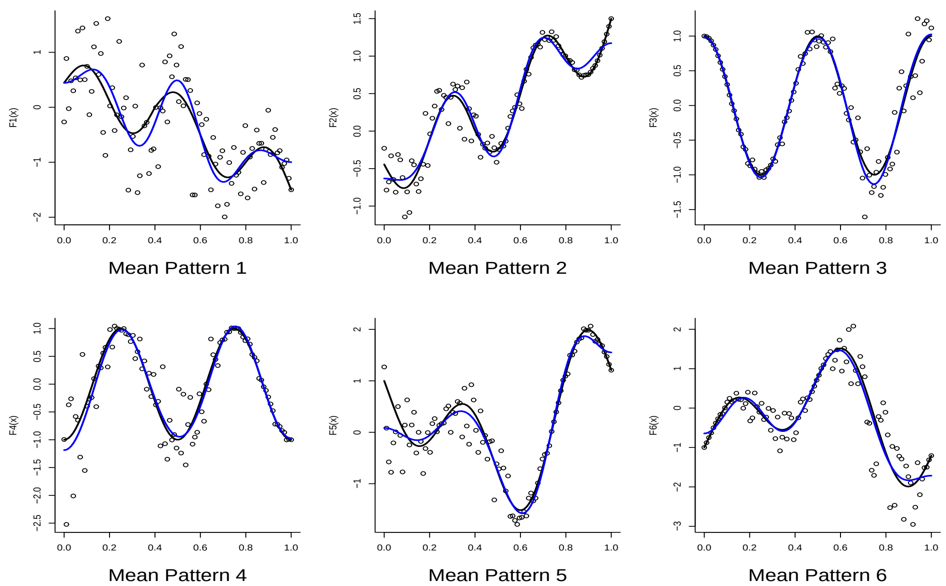

Figure 4: Six synthetic curves, their true mean patterns (in black) and their estimated mean pattern (in blue).

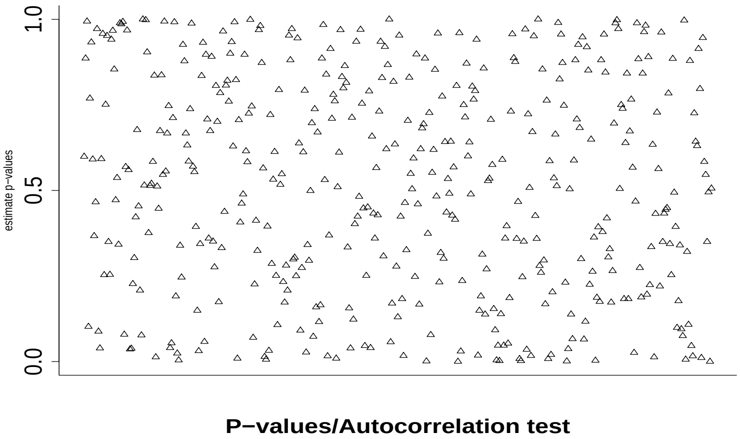

Figure 5: Estimated p-values for autocorrelation test for research and development expenditure data.

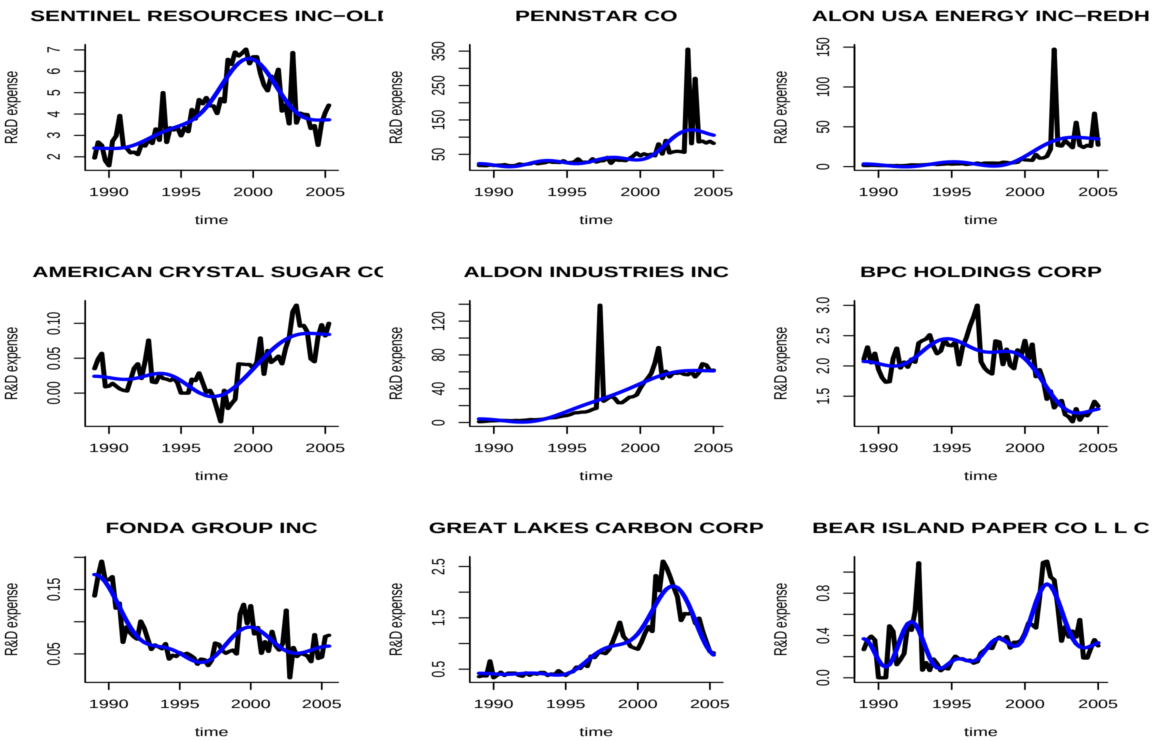

Figure 6: Nine randomly chosen curves in the R&D expenditure data and their estimated patterns (in blue).

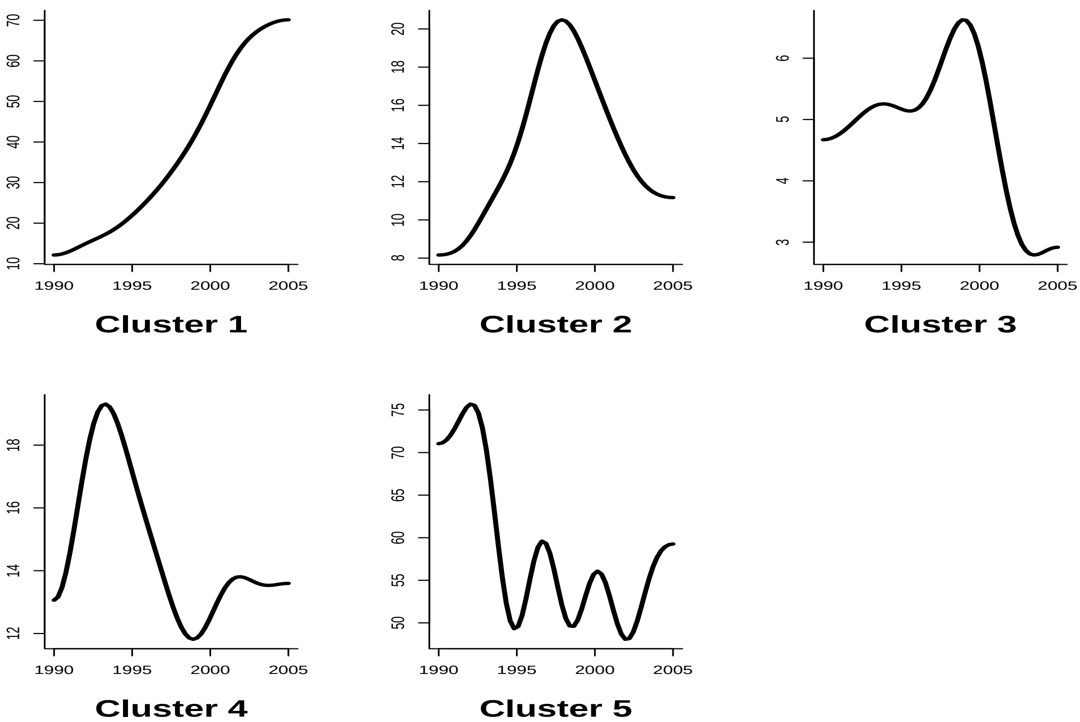

Figure 7: Mean profiles of five clustered of the R&D expenditure curves.