Clustering Seasonality Patterns in the Presence of Errors

Mahesh Kumar
Operations Research Center
MIT
Cambridge, MA 02139
maheshk@mit.edu

Nitin R. Patel
Sloan School of Management
MIT
Cambridge, MA 02139
nitinrp@mit.edu

Jonathan Woo
ProfitLogic Inc.
Cambridge, MA 02142
jwoo@profitlogic.com

ABSTRACT

Clustering is a very well studied problem that attempts to group similar data points. Most traditional clustering algorithms assume that the data is provided without measurement error. Often, however, real world data sets have such errors and one can obtain estimates of these errors. We present a clustering method that incorporates information contained in these error estimates. We present a new distance function that is based on the distribution of errors in data. Using a Gaussian model for errors, the distance function follows a Chi-Square distribution and is easy to compute. This distance function is used in hierarchical clustering to discover meaningful clusters. The distance function is scale-invariant so that clustering results are independent of units of measuring data. In the special case when the error distribution is the same for each attribute of data points, the rank order of pair-wise distances is the same for our distance function and the Euclidean distance function. The clustering method is applied to the seasonality estimation problem and experimental results are presented for the retail industry data as well as for simulated data, where it outperforms classical clustering methods.

Keywords

clustering, distance function, forecasting, Gaussian distribution, product life cycle, seasonality, time-series.

1. INTRODUCTION

Definition of a good distance or dissimilarity function is a critical step in any distance based clustering method. Most of the work in this field assumes that a distance function defined on pairs of objects is available [3, 4]. Sometimes these distances are directly measured as pair-wise differences among objects, but often they are computed from measurements of a number of attributes for each object. Most traditional clustering methods assume that these distances (or dissimilarities) can be computed from the data in hand without any error. However, in certain applications such as the seasonality estimation problem described below, data representing each object is not observable. A statistical method is applied to estimate the data. For example, if one wishes to cluster various geographical regions based on per household income and expenditure, one might represent each geographical region by the average household income and average expenditure. A sample average by itself is inadequate and can be misleading unless the variation around the average is negligible. In some applications (e.g., random sampling) it is easy to obtain an estimate of the deviation along with the average. These deviations, which are measures of error for the averages, may be very different for different data points. In this paper we present a method of clustering using information about the errors in data. Although our study and results are focused on time-series clustering in the retail industry, the concept can be extended to other clustering applications where measurement errors are significant.

Numerous approaches to clustering include statistical, machine learning and optimization perspectives [2, 3, 4]. To the best of our knowledge, all these approaches assume that each data point is observed without any error. The fundamental difference of our approach is that we model measurement errors.

We assume that we are given n points in the T-dimensional real space and there is an observation error associated with each data point. The observation errors may come from different distributions. We assume that we are given a value and a standard error for each data point. Our goal is to group these data points into clusters so that it is very likely that points in the same cluster have the same true value whereas it is quite unlikely for points in different clusters to have the same true value (likelihood being defined with respect to the probability distributions of the errors associated with the data points). In contrast to standard clustering solutions, in this case, two points that differ significantly in their observed values might belong to the same cluster if they have high errors associated with them whereas two points that do not differ much in their observed values might belong to different clusters if they have small errors. The above argument is illustrated in Figure 1. Four points A, B, C and D are the observed values of four objects and these values have Gaussian errors associated with them as represented by the ellipse surrounding each data point. Clustering that recognizes errors will put points A and B into one cluster and points C and D in another, whereas a clustering method that does not consider errors will cluster A and C together and B and D together. In this example the clustering result from error-based clustering makes

KDD '02 Edmonton, Alberta, Canada

more sense because the x values have large error in their measurements, whereas, y value measurements are accurate and should therefore dominate the clustering decision.

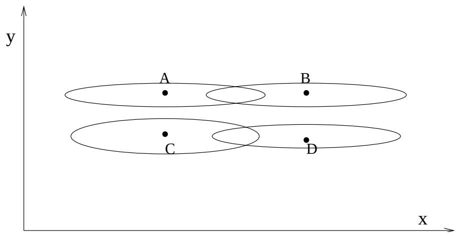

Figure 1: Data points along with errors

Our research was motivated by the problem of estimating seasonality for retailers based on the sales data from the previous year. In retail merchandizing it is very important to understand the seasonal behavior in the sales of different items to correctly forecast demand and make appropriate business decisions. We model the seasonality estimation problem as a time-series clustering problem in the presence of errors and present the experimental results when applied to point-of-sale retail data. We were able to discover meaningful clusters of seasonality whereas classical methods which do not take account of errors did not obtain good clusters. To the best of our knowledge, this is not only the first attempt to cluster data while incorporating information about errors in data, we have not come across any work that attempts to find seasonal patterns in retail marketing using time-series clustering.

Although our studies and results can be extended to arbitrary probability distributions, we assume that each point comes from a multidimensional Gaussian distribution with diagonal covariance matrix since this distribution is appropriate for our application. Under this assumption we present a new distance function that is based on the distribution of error in data. Under a Gaussian model for errors, the distance function follows a Chi-Square distribution and is easy to compute. The distance function is used in hierarchical clustering to develop a clustering method that is used in the seasonality estimation problem.

The rest of the paper is organized as follows. In section 2 we briefly describe the seasonality estimation problem. In section 3 we provide a time-series representation of seasonality that models associated errors. In section 4 we define a distance function based on the distribution of errors in data. In section 5 we describe a hierarchical clustering algorithm using this distance function. In section 6 we present experimental results based on real as well as simulated data. Finally in section 7 we present concluding remarks along with future research directions.

2. SEASONALITY ESTIMATION

Seasonality is defined as the normalized underlying demand of a group of similar merchandize as a function of time of the year *after taking into account other factors that impact sales such as discounts, inventory, promotions and random effects.* Seasonality is a numeric index of seasonal buying behavior that is consistent from year to year. For example, a Christmas item will have high seasonality indices during the month of December, whereas shorts will have consistently high seasonality indices during summer and low indices during winter. There are many different possible seasonal patterns (see figures 6 & 7). In the retail industry, practical concerns regarding logistic complexity require that we handle no more than a few (5-15) seasonal patterns. Therefore, our goal is to identify a small set of seasonal patterns that model the items sold by the retailer and relate each item to one of these seasonal patterns.

Considerable work has been done on how to account for the effect of price, promotions inventory and random effects[6, 8]. In our retail application, weekly sales of an item i are modelled as products of several factors that affect sales as described in equation (1).

$$Sale_{it} = f_I(I_{it}) * f_P(P_{it}) * f_Q(Q_{it}) * f_R(R_{it}) * PLC_i(t-t_0^i) * Seas_{it} \tag{1}$$

Here, I_{it}, P_{it}, Q_{it} and R_{it} are the quantitative measures of inventory, price, promotion and random effect respectively for an item i during week t. f_I, f_P, f_Q and f_R model the impact of inventory, price, promotion and random effect on sales respectively. PLC is the Product Life Cycle coefficient which is defined as the sale of an item in the absence of seasonality as well as the factors discussed above. The shape and duration of the PLC curve depends on the nature of the item. For example, a fashion item will sell out very fast compared to a non-fashion item as shown in the figure 2. $Sale_{it}, Seas_{it}$ and $PLC_i(t-t_0^i)$ are the sale value, seasonality coefficient and PLC coefficient of item i during week t where t_0^i is the week when the item is introduced. For convenience we define the PLC value to be zero during weeks before the item is introduced and after it is removed. Seasonality coefficients are relative. To compare seasonality coefficients of different items on the same scale, we assume that sum of all seasonality coefficients for an item over a year is constant, say equal to the total number of weeks, which is 52 in this case.

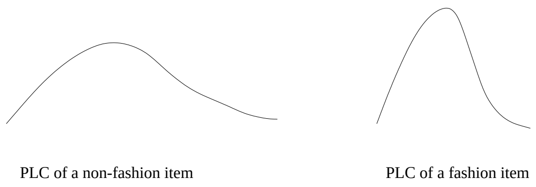

Figure 2: PLCs for non-fashion and fashion items

In this paper we will assume that our data has been preprocessed by using (1) to remove the effects of all these non-seasonal factors. We also assume that the data has been normalized to enable comparison of sales of different items on the same scale. After pre-processing and normalization the sale of an item, $Sale_{it}$ is determined by the PLC and seasonality as described below.

$$Sale_{it} = PLC_i(t - t_0^i) * Seas_{it}. \tag{2}$$

Since adjusted sales of an item is the product of PLC and seasonality, it is not possible to determine seasonality just by looking at the sale values of an item. The fact that items having the same seasonality pattern might have different PLCs complicates the analysis.

Initially, based on domain knowledge from merchants we group items that are believed to follow similar seasonality over the entire year. For example, one could group together a specific set of items that are known to be selling during Christmas, all items that are known to be selling during summer and not during winter, etc. The idea is to get a set of items following similar seasonality that are introduced and removed at different points of time during the year. This set, say S, consists of items having a variety of PLCs differing in their shape and time duration. If we take the weekly average of all PLCs in S then we would have a somewhat flat curve as shown in figure 3. This implies that weekly average of PLCs for all items in S can be assumed to be constant as shown in theorem 1.

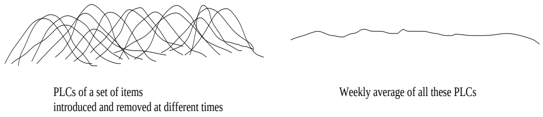

Figure 3: Averaging effect on a set of uniformly distributed PLCs

Theorem 1. *For a large number of PLCs that have their introduction dates uniformly spread over different weeks of year, the weekly average of PLCs is approximately constant, i.e.,*

$$\frac{1}{|S|}\sum_{i\in S} PLC_i(t-t_0^i) \approx c \qquad \forall t = 1,...,52 \tag{3}$$

Proof. Let us consider a given week, say week t. Since only those PLCs that have starting time between week $t-51$ and week t will contribute to the weekly average for week t, we consider only those PLCs that have t_0^i between week $t-51$ and week t. Let p_l be the probability of $t_0^i = l$. Because of equally likely starting times, $p_l = \frac{1}{52}$ for $l = t-51, t-50, ..., t$.

$$E\Big(\frac{1}{|S|}\sum_{i\in S} PLC_i(t-t_0^i)\Big) = \frac{1}{|S|}\sum_{i\in S} E(PLC_i(t-t_0^i))$$

$$= \frac{1}{|S|}\sum_{i\in S}\sum_{l=t-51}^{t} p_l * PLC_i(t-l) = \frac{1}{52*|S|}\sum_{i\in S}\sum_{l=0}^{51} PLC_i(l) = c \tag{4}$$

where c is a constant that does not depend on t. The variance of $\frac{1}{|S|}\sum_{i\in S} PLC_i(t-t_0^i)$ is inversely proportional to $|S|$ as in equation (8). If $|S|$ is large, the variance will be small and the weekly observed values of $\frac{1}{|S|}\sum_{i\in S} PLC_i(t-t_0^i)$ will be approximately constant and hence the result. □

If we take the average of weekly sales of all items in S then it would nullify the effect of PLCs as suggested by the following equations. Let $Sale_t$ be the average sale during week t for items in S then

$$Sale_t = \frac{1}{|S|}\sum_{i\in S} Sale_{it} = \frac{1}{|S|}\sum_{i\in S} Seas_{it} * PLC_i(t-t_0^i). \tag{5}$$

Since all items in S are assumed to have the same seasonality, $Seas_{it}$ is the same for all items $i \in S$, say equal to $Seas_t$, i.e.,

$$Seas_{it} = Seas_t \qquad \forall i \in S, \qquad t = 1,2,..,52. \tag{6}$$

Therefore,

$$Sale_t = Seas_t * \frac{1}{|S|}\sum_{i\in S} PLC_i(t-t_0^i) \approx Seas_t * c \quad t = 1,...,52. \tag{7}$$

The last equality follows from theorem 1. Thus seasonality values, $Seas_t$, can be estimated by appropriate scaling of weekly sales average, $Sale_t$.

The average values obtained above will have errors associated with them. An estimate of the standard error in $Sale_t$ is given by the following equation.

$$\sigma_t = \sqrt{\frac{1}{|S|}\sum_{i\in S}(Sale_{it} - Sale_t)^2} \tag{8}$$

The above procedure provides us with a large number of seasonal patterns, one for each set S, along with estimates of associated errors. The goal is to form clusters of these seasonal patterns based on their average values and errors. Each cluster of seasonal patterns is finally used to estimate seasonality of the cluster. This estimate will have smaller error than if we estimated seasonality for each pattern in S.

One might attempt to estimate seasonality using standard time-series clustering, but the danger of not incorporating knowledge of errors in the clustering method is that the information on variability of the data points is ignored. Knowledge of errors would lead us to be more careful in assigning a low error data point to a cluster than a high error data point. Errors and associated probability distributions capture the variability of a data point and, in the rest of the paper, we will present how explicit treatment of errors can be used to discover better clusters.

3. REPRESENTATION OF TIME-SERIES

Time-series data differs from other data representations in that a data point in time-series is represented by a sequence typically measured at equal time intervals. Various time-series representations have been proposed in [1, 5] for data with no errors. In this section we present a time series representation that models errors associated with data.

In our model a time-series sampled at T points is represented by a sequence of T distributions. We assume that each of these T samples are independent of each other and are distributed according to one-dimensional Gaussian distributions. A time-series is represented as $A = \{(x_1,\sigma_1), (x_2,\sigma_2), ..., (x_T,\sigma_T)\}$ where the t^{th} sample of A is normally distributed with mean x_t and standard deviation σ_t.

We assume that all the T samples of a time-series are normalized, i.e., $\sum_t x_t = T$. For seasonality estimation we have T equal to 52 corresponding to the number of weeks in a year. x_t is the normalized value of seasonality estimate, $Sale_t$, obtained in equation (7). σ_t is the standard error in the estimated value of x_t as in equation (8).

4. DISTANCE FUNCTION

Like most clustering methods we assume that the relationships among a set of n objects is described by an $n \times n$ matrix containing a measure of dissimilarity between the i^{th} and the j^{th} data points. In clustering parlance it is referred to as the distance function between a pair of points. Various distance functions have been considered for the setting where data has no measurement errors [3, 4]. In this section we develop a probability based distance function for data with errors.

4.1 Distance Function Definition

Consider two seasonalities $A_i = \{(x_{i1}, \sigma_{i1}), (x_{i2}, \sigma_{i2}), ..., (x_{iT}, \sigma_{iT})\}$ and $A_j = \{(x_{j1}, \sigma_{j1}), (x_{j2}, \sigma_{j2}), ..., (x_{jT}, \sigma_{jT})\}$. A_i and A_j are the estimated values of two seasonalities as described in section 2. Let the corresponding true seasonalities be $\{\mu_{i1}, \mu_{i2}, ..., \mu_{iT}\}$ and $\{\mu_{j1}, \mu_{j2}, ..., \mu_{jT}\}$. This means that x's are the observed values that come from distributions with true means of μ's. We define similarity between two seasonalities as follows. If the null hypothesis $H_0 : A_i \sim A_j$ is true then similarity between A_i and A_j is the probability of accepting the hypothesis. Here, $A_i \sim A_j$ denotes $\mu_{it} = \mu_{jt}$ for $t = 1, ..., T$. The distance d_{ij} between A_i and A_j is defined as (1-similarity), which is the probability of rejecting the above hypothesis. This distance function satisfies the following desirable properties.
$dist(A, B) = dist(B, A)$
$dist(A, B) \geq 0$
$dist(A, A) = 0$
$dist(A, B) = 0 \Leftrightarrow A = B$
$dist(A, B) \leq 1$.

Let $N(\mu, \sigma)$ denote a Gaussian distribution with mean μ and standard deviation σ. Consider t^{th} samples of both seasonalities, $A_{it} = (x_{it}, \sigma_{it})$ and $A_{jl} = (x_{jl}, \sigma_{jl})$. We assumed that A_{it} and A_{jt} come from independent Gaussian distributions with means μ_{it} and μ_{jt} and standard deviations σ_{it} and σ_{jt} respectively. This implies that $x_{it} \sim N(\mu_{it}, \sigma_{it})$ and $x_{jt} \sim N(\mu_{jt}, \sigma_{jt})$ so that $(x_{it} - x_{jt}) \sim N(\mu_{it} - \mu_{jt}, \sqrt{\sigma_{it}^2 + \sigma_{jt}^2})$.

If $A_i \sim A_j$ then $\mu_{it} = \mu_{jt}$ and consequently the statistic $\frac{x_{it} - x_{jt}}{\sqrt{\sigma_{it}^2 + \sigma_{jt}^2}}$ follows a t-distribution, but if a large amount of data is used in the estimation of x_{it} and x_{jt} then it can be approximated by the standard Gaussian distribution $N(0, 1)$ [7]. Therefore, $\frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2}$ will follow the Chi-Square distribution with one degree of freedom. Since the T samples are assumed to be independent the statistic $\sum_{t=1}^{T} \frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2}$ follows the Chi-Square distribution with $T - 1$ degrees of freedom (1 degree less because of the constraint: $\sum_{t=1}^{T} x_{it} = \sum_{t=1}^{T} x_{jt}$). Therefore, $1 - \chi_{T-1}^2(\frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2})$ is the probability of accepting two seasonalities as the same in spite of having the observed differences. Consequently,

$$d_{ij} = \chi_{T-1}^2(\frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2}). \tag{9}$$

4.2 Comparison with Euclidean Distance

Since $\chi_f^2(x)$ is a monotonically increasing function w.r.t. x for any degrees of freedom f, d_{ij} is monotonically increasing with respect to $\sum_{t=1}^{T} \frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2}$. Therefore, among all pairwise distances between the given time-series sequences, the rank order of d_{ij} is the same as that of $\frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2}$. If all σ's were the same and equal to σ then it would become the rank order of $\frac{1}{2\sigma^2} \sum_{t=1}^{T} (x_{it} - x_{jt})^2$ which is the same as the rank order of the Euclidean distance, $\sum_{t=1}^{T} (x_{it} - x_{jt})^2$. Therefore, when all the errors are equal, the proposed distance function has the same rank order as the Euclidean distance. Popular distance-based hierarchical clustering methods use single linkage and complete linkage. For these methods it is the rank order of distances that matters and not the actual distances. Therefore, a clustering method based on the proposed distance function will be identical to a clustering method based on $\sum_{t=1}^{T} \frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2}$ which is a weighted Euclidean distance function where each sample is weighted with the inverse of its pooled error. This makes intuitive sense because it gives smaller weight to the data that have higher error and large weight to samples that have small error.

4.3 Scale Invariant Clustering

Many distance functions used in clustering change nonlinearly with change in scale of measuring data and subsequently the clustering results might also change [3]. The distance function we have proposed is independent of scale. When we change units of measurement of data, the observed x values and corresponding errors σ's are multiplied by the same factor. Therefore, $\sum_{t=1}^{T} \frac{(x_{it} - x_{jt})^2}{\sigma_{it}^2 + \sigma_{jt}^2}$ is unit-free and so d_{ij} is scale invariant.

5. CLUSTERING

5.1 Clustering Algorithm

Definition of a good distance function is the most critical step in any distance based clustering method. Having decided on a distance function we use a hierarchical clustering method that is similar to Ward's method [9]. In this method, we start with each data point being a singleton cluster and at each stage combine the closest pair of clusters into a single cluster until a threshold value is reached or a predefined number of clusters is obtained. At each intermediate stage two clusters are combined into a single cluster using a *merge* operation defined in section 5.2. The clustering algorithm is formally described below.

```
  Algorithm $hError(A, G)$
Input: $A_i = \{(x_{i1}, \sigma_{i1}), (x_{i2}, \sigma_{i2}), ..., (x_{iT}, \sigma_{iT})\}, i = 1, 2, ...n$
         $G$ = number of clusters.
Output: $Cluster(i), i = 1, 2, ..., G$.
Start
  for $i = 1$ to $n$
    $Cluster(i) = \{i\}$
    $seas(i) = A_i$
  end
  $NumClust = n$
  while $NumClust > G$
    for $1 \leq i < j \leq NumClust$
      calculate $d_{ij} = dist(seas(i), seas(j))$ using
                equation (9)
    end
    $(I, J) = \arg\min_{1 \leq i < j \leq NumClust} d_{ij}$
    $Cluster(I) = Cluster(I) \cup Cluster(J)$
    $seas(I) = merge(seas(I), seas(J))$ using
           equations (10) and (11)
    $Cluster(J) = Cluster(NumClust)$
    $seas(J) = seas(NumClust)$
    $NumClust = NumClust - 1$
  end
  return $Cluster(i), i = 1, 2, ..., G$
end
```

5.2 Merging time-series

The *merge* operation is used in the Algorithm *hError* to combine information from a pair of time-series to produce a new time-series that is an interpolation between the time-series used to produce it. The shape of the resulting time-series depends not only on the sample values of individual time-series but also on errors associated with them.

Consider two time-series, $A = \{(x_{11}, \sigma_{11}), (x_{12}, \sigma_{12}), ..., (x_{1T}, \sigma_{1T})\}$ and $B = \{(x_{21}, \sigma_{21}), (x_{22}, \sigma_{22}), ..., (x_{2T}, \sigma_{2T})\}$. Let $C = \{(x_1, \sigma_1), (x_2, \sigma_2), ..., (x_T, \sigma_T)\}$ be the resulting time-series when A and B are merged. Let A and B come from the same true seasonality with means $\{\mu_1, \mu_2, ..., \mu_T\}$. A natural choice for the components of C are the maximum likelihood estimates of μ's and associated standard deviations. From the maximum likelihood principle [7] and the Gaussian distribution assumption, it is easy to show that

$$x_t = \frac{1}{\frac{1}{\sigma_{1t}^2} + \frac{1}{\sigma_{2t}^2}} \left(\frac{x_{1t}}{\sigma_{1t}^2} + \frac{x_{2t}}{\sigma_{2t}^2}\right) \qquad t = 1, 2, ..., T. \qquad (10)$$

$$\sigma_t = \frac{1}{\sqrt{\frac{1}{\sigma_{1t}^2} + \frac{1}{\sigma_{2t}^2}}} \qquad t = 1, 2, ..., T. \qquad (11)$$

6. EXPERIMENTAL RESULTS

In this section we present experimental results using Algorithm *hError* on simulated data and also on data from a leading national retail chain.

6.1 Simulated Data

We generated artificial data using ten PLCs that differ in their peaks and shapes as shown in figure 4. The PLC data is randomly generated by choosing one of these ten PLCs with equal probability and a uniformly distributed starting time over a period of one year. Using three different seasonalities corresponding to Christmas, summer seasonality and winter seasonality (see figures 6 & 7), we generated sales data by multiplying each generated PLC with one of the three seasonalities. We constructed 12 instances, where each instance consists of 25-35 PLCs. Sales data for each instance was generated by multiplying all the PLCs in that instance with one of the above seasonalities chosen at random. We hide the information about true seasonalities and use Algorithm *hError* to recover three seasonalities.

We obtain an estimate of seasonality and associated errors for each instance by averaging weekly sales data in that instance as described in section 2. The estimated seasonalities and associated errors are shown in figure 5 with vertical bars representing standard errors. It can be seen from the figure that some of the seasonalities do not correspond to any of the original seasonalities, for example, the middle one in the last row. Moreover, each of them has large errors. We ran *hError* and obtained the three cluster centers shown in figure 6. The resulting seasonalities match original seasonalities very well as can be seen from this figure. We compared our result against k-means and Ward's method that do not consider the information about errors. The number of misclassifications were higher when we used these clustering methods. The clusters were identical for both of them and the cluster centers are shown in figure 7.

We assess the quality of a clustering result by computing its *Average Estimation Error*. Let there be r true seasonalities, $Seas_1$, $Seas_2$, ..., $Seas_r$. r is 3 in the above experiment. Let the estimated seasonalities be $Estimate_1$, $Estimate_2$, ..., $Estimate_r$ then the *Average Estimation Error* is defined as

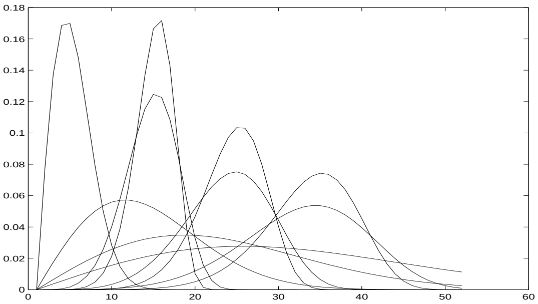

Figure 4: Ten different PLCs.

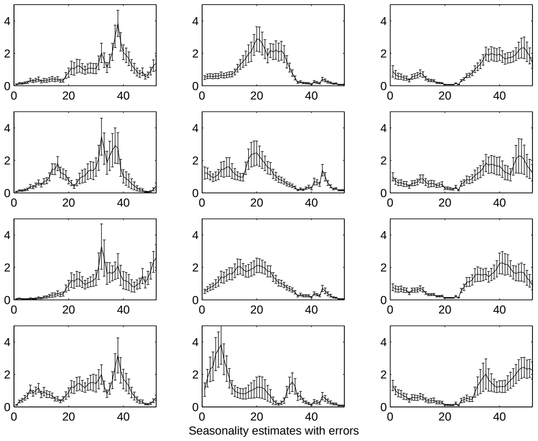

Figure 5: Individual (prior to clustering) seasonality estimates with associated errors.

$$AverageEstimationError = \frac{1}{r} \sum_{i=1}^{r} |Seas_i - Estimate_i| \qquad (12)$$

where $|Seas_i - Estimate_i|$ is the total absolute difference between the true seasonality indices and the estimated seasonality indices defined as below.

$$|Seas_i - Estimate_i| = \sum_{t=1}^{52} |Seas_{it} - Estimate_{it}| \qquad (13)$$

In the above experiment the *Average Estimation Error* was 4.9758 using kmeans or Ward's method, whereas it was only 1.7780 using *hError*.

We replicated the above experiment 100 times. Table 1 shows the average number of misclassifications and *Average Estimation Error* made by different clustering methods on a set of 12 seasonalities when clustered into 3 groups.

6.2 Retailer Data

In order to investigate the usefulness of our technique in practice, we carried out comparative analysis on real data from a major retail chain. Retail merchandize is divided into several departments (examples: shoes, shirts, jewelry.)

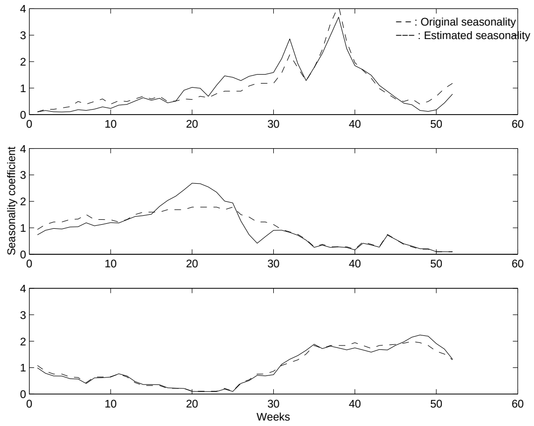

Figure 6: Seasonalities obtained by *hError*.

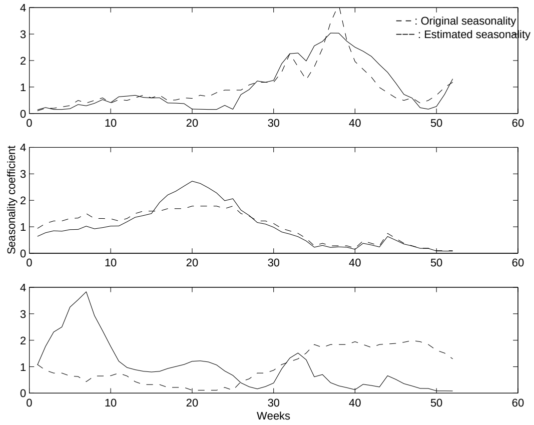

Figure 7: Seasonalities obtained by kmeans and Ward's method using Euclidean distances.

which are further classified into several classes (example: men's winter shoes, formal shirts, etc.). Each class has a varying number of items for which sales data is available. For this experiment we considered only those classes that have sales data for at least 20 items. The data used consisted of two years of sales data. One year of data was used to estimate seasonalities. Using these estimated seasonalities, we forecast sales for the next year and compare it against the actual sales data. We considered 6 different departments (greeting cards, books, music and video, toys, automotive, and sporting goods). Each department has 4-15 classes and we used data from a total of 45 classes across all 6 departments. First we estimated seasonalities and associated errors for each class based on the method described in section 2. Having estimated seasonalities, we applied Algorithm *hError* to reconstruct seasonalities for each class. Using these seasonality estimates, we predicted sales for the items in the books department. We chose the books department because the effects such as price, promotions and inventory were small for this department, thereby, weekly change in sales for the books department was mainly because of seasonality. We assessed the quality of forecast by calculating average *Forecast Error*, which is the ratio of the total difference between actual sale and forecast sale to the total actual sale, as defined below.

Table 1: Average # misclassifications and Average Estimation Error for different clustering methods.

Clustering Method	Average # misclassifications	Average Estimation Error
hError	0.87	2.0182
Ward's method	2.63	4.7021
kmeans	2.94	5.0337

$$ForecastError = \frac{\sum_{t=1}^{T} |ActualSale_t - ForecastSale_t|}{\sum_{t=1}^{T} ActualSale_t} \quad (14)$$

We compared our result against kmeans and Ward's method based on the Euclidean distance. We also compared our forecast when no clustering was used, i.e., when the forecast was based on the seasonality estimates for each class using average of weekly sales data as described in section 2. We found that forecasts using *hError* were substantially better than forecasts using kmeans or Ward's method or forecasts without using clustering. Table 2 compares average *Forecast Error* in these four situations for 17 different items in the books department.

Table 2: Average Forecast Error

Clustering Method	Average Forecast Error %
hError	18.7
Ward's	23.9
kmeans	24.2
No clustering	31.5

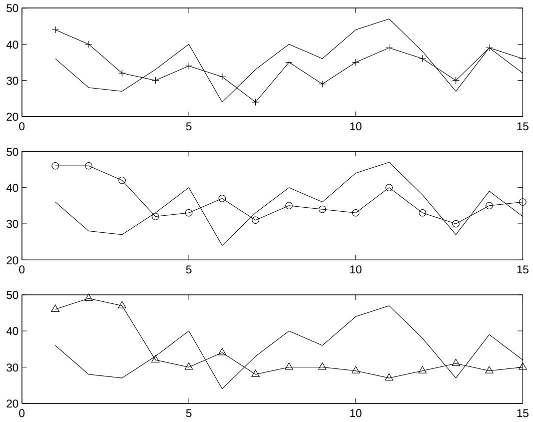

Figure 8: Sales forecast against actual sales
——— : actual sales
—+— : forecast using *hError*
—o—- : forecast using Ward's method
—△— : forecast without clustering

Figure 8 shows graphs comparing these forecasts for one item in the books department. This item was sold for a total of 33 weeks during January through September 1997.

The price for the item was constant during this period and there was no promotion on this item, therefore we ignored all external factors and made our forecast using only PLC and seasonality coefficients. Seasonality of the class that contains this item is estimated using past year's sales data of all the items in the class. The first 18 weeks of sales data of this item is used to estimate the PLC. PLC is estimated by simple curve fitting from a set of predefined PLCs. Using the seasonality estimates from past year's data and PLC estimate from the first 18 weeks of data, we forecast sales for the remaining 15 weeks. The graphs show that forecasts using *hError* are significantly better than the others.

In figure 8 we observe that seasonality estimates without clustering failed to capture the sales pattern. Clustering using Euclidean distance succeeded in making a better forecast but clustering using error-based distance function was even better. The reason is that the books department has 5 classes. Because very few items are used to estimate seasonality for each class, each seasonality estimate has large errors and therefore the forecast based on this seasonality estimate (without clustering) does not match actual sales. On close inspection of the data we found that there are two groups of 3 and 2 classes having similar seasonalities. Clustering identifies the right clusters of 3 and 2 seasonalities. The combined seasonality of each cluster has higher accuracy because more items are used to estimate it. Clustering using the error-based distance function does better than Euclidean distance based clustering because it gives more weight to seasonality with smaller errors obtained by using larger number of items.

We restricted our forecast analysis to only a small section of books items that had small fluctuation in price over their selling period. This helped us eliminate effects due errors in estimation of the factors relating to price or promotion.

7. DISCUSSION AND FUTURE RESEARCH

In this paper we have developed a clustering method that incorporates information about errors associated with data. Traditional clustering methods are inadequate when different data points have very different errors. We introduced a new distance function which is based on Gaussian distribution of errors. We showed that this distance function can be viewed as a generalization of the classical Euclidean distance. The distance function also has the property that it is invariant under different scales for data. Finally, we demonstrated the utility of our method, on both simulated and real data sets, in improving estimates of seasonality in the retail industry.

Although we developed the distance function for time-series clustering, the concept of incorporating information about error in the distance function is very general and can be used in many other clustering applications. In our research we made a basic assumption that the T samples of a time-series come from independent distributions. In time-series data, it is common to encounter positive correlations in consecutive sample values. Less often we also encounter negative correlations. Therefore, while incorporating the concept of dependence can be difficult, it can improve the test statistic developed in section 4 and subsequently give more accurate measure of the distance function. We have generalized our approach to accomodate correlated observations. We have a working paper under development with this theme. In the case of seasonality estimation problem we deal with seasonality values that are obtained by taking average of sales data of a group of items. Dependency among samples of a time-series might not be a serious problem for seasonality estimation because the averaging process might dampen the effect of dependency.

In this paper we have provided a hierarchical clustering heuristic that recognizes errors associated with data. The next step would be to formulate a model for the problem and let the model provide a basis for a natural heuristic to solve the problem. We have already developed a model that justifies the work presented here and extended it to correlated data. We are in the process of developing and comparing several heuristics that arise from our model. We are also exploring other applications of error-based clustering. We have already identified that the method works very well in clustering of regression coefficients. We expect to be able to report this work in progress soon.

Errors are natural in any data measurement. Often errors contain very useful information and should be considered an important part of data. We feel that a clustering method using the information contained in errors is an important conceptual step in the field of cluster analysis and data mining.

8. ACKNOWLEDGEMENT

This work was supported by ProfitLogic Inc. and the e-business Center at Sloan MIT. The authors gratefully acknowledge the contributions of James B. Orlin in this work.

9. REFERENCES

[1] Rakesh Agrawal, King-Ip Lin, Harpreet S. Sawhney, Kyuseok Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. *VLDB*, 490-501, 1995.

[2] Scott Gafney, Padhraic Smyth. Trajectory clustering with mixtures of regression models. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999.

[3] A. K. Jain, R. C. Dubes. Algorithms for Clustering Data. *Prentice-Hall*, 1988.

[4] A. K. Jain, M. N. Murty, P. J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, Volume 31, No. 3, 264-323, 1999.

[5] Eamonn J. Keogh, Michael J. Pazzani. An enhanced representation of time-series which allows fast and accurate classification, clustering and relevance feedback. *Fourth conference on Knowledge Discovery in Databases and Data Mining*, 1998.

[6] Praveen K. Kopalle, Carl F. Mela, Lawrence Marsh. The dynamic effect of discounting on sales: Empirical analysis and normative pricing implications. *Marketing Science* , 317-332, 1999.

[7] John A. Rice. Mathematical Statistics and Data Analysis. Second Edition. *Duxbury Press.*

[8] Jorge M. Silva-Risso, Randolph E. Bucklin, Donald G. Morrison. A decision support system for planning manufacturers' sales promotion calendars. *Marketing Science*, 274-300, 1999.

[9] J. H. Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* Volume 58, Issue 301, 236-244, 1963.