Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 108C (2017) 2079–2089

Procedia Computer Science

International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland

# Optimal pricing model based on reduction dimension: A case of study for convenience stores

Laura Hervert-Escobar, Oscar A. Esquivel-Flores, and Raul V. Ramirez-Velarde

Instituto Tecnológico y de Estudios Superiores de Monterrey. Monterrey, Nuevo León, México

**Abstract**

This paper proposes a methodology to define an optimal pricing strategy for convenience stores based on dimension reduction methods and uncertainty of data. The solution approach involves a multiple linear regression (MLR) as well as a linear programming optimization model. Two strategies Principal Component Analysis (PCA) and Best Subset Regression (BSR) methods for the selection of a set of variables among a large number of predictors is presented. A linear optimization model then is solved using diverse business rules. To show the value of the proposed methodology optimal prices calculation results are compared with previous results obtained in a pilot performed for selected stores. This strategy provides an alternative solution that shows how a decision maker can include proper business rules of their particular environment in order to define a pricing strategy that meets business goals.


Peer-review under responsibility of the scientific committee of the International Conference on Computational Science

*Keywords:* Dimension reduction, Multiple Linear Regression, Optimization, Optimal Pricing

## 1 Introduction

A convenience store is a small retail business that stocks a range of everyday items such as groceries, snack foods, candy, toiletries, soft drinks, tobacco products, magazines and newspapers. Convenience stores typically have longer opening hours, serve more locations, and have shorter cashier lines. Therefore, it is thought that such kind of store charge significantly higher prices than conventional grocery stores or supermarkets, as convenience stores order smaller quantities of inventory at higher per-unit prices from wholesalers. However, it is important to find a correct level of prices in the correct level in order to keep the customers and increase their profits. Therefore, a price strategy in this business is relevant because the price should increase the profits but also keep customers happy.

Research pricing methods fall into two main approaches: pricing the total product (pricing for a complete concept/product) or pricing other elements such as branding or features.

According to [6], the pricing strategy is by definition the effort aimed at finding a product's optimum price, typically including overall marketing objectives, consumer demand, product attributes, competitors' pricing, and market and economic trends.

1877-0509 
Peer-review under responsibility of the scientific committee of the International Conference on Computational Science
10.1016/j.procs.2017.05.269

Applied research data analysis and forecasts have been treated using linear regression [14, 1] considering multiple explanatory variables combined with other methods for selecting variables in order to reduce the dimensionality of the problem, for instance stepwise and genetic algorithms [12].

Multiple linear regressions (MLR) model is one of the modelling techniques to investigate the relationship between a dependent variable and several independent variables.

MLR is a widely used technique to refine decision making in finance and economics. Ismail, et al. [7] use MLR as a tool to develop forecasting model for predicting gold prices based on economic factors. Shabri and Samsudin[2] propose a hybrid model which integrates wavelet regression and multiple linear regression for crude oil price forecasting, it is important to highlight the use of principal component analysis (PCA) to process data subs-series in MLR to reduce the dimensions of sub-time

Dimension reduction problem has been treated widely, Chun and Keleş [4] propose a sparse partial least squares formulation which aims to achieve variable selection by producing sparse linear combinations of the original predictors. Li-Ping et al [16] offer a methodology to sufficient dimension reduction and propose alternating inverse regression to estimate the central subspace which circumvents the collinearity and curse of dimensionality simultaneously [15].

The problem definition of this research considers a company that operates several convenience stores. The company desires to improve the current pricing strategy in order to maximize the business performance. The main goals are to obtain the optimum price of each product and to increase the gross margin. A first approach presented in [5] considers a three-stage proposal. First stage consists in reducing variables in order to keep the most appropriate variables. This reduction is based on Pearson's correlations and the use of a ratio based on interaction of independent vs explanatory variables. The ratio is helpful to reduce the multicollinearity issues. Then an econometric model was obtained through MLR methodology using a stepwise algorithm for reducing dimensionality, and finally the optimization model with additional business rules (internal and external) was performed in order to obtain an optimal price proposal. The strategy was tested in a real case study, where the results proved the value of the methodology.

This research is an extension of the first approach that improves each step. In this way, the first stage consists on the selection of significant variables using two methods in parallel (in order to compare later). The methods are PCA and BSR algorithms, then a MLR is performed. The multicollinearity issues are resolved using Ridge Regression. Finally, the optimization model is improved by introducing uncertainty in the expected volume sales. This variable was not introduced in the first approach.

The organization of the document is as follows, the introduction above served to set the business frame as well as to present the literature review and define problem. Then, the mathematical formulation of the optimization model is presented in section 2. Section 3 describes the design and implementation of the methodology. Section 4 compares the results with a previous methodology. The last section 5 consists of the conclusion and outlines the future work.

# 2 Mathematical Formulation of the optimization model

The pricing strategy proposed in this research is structured in three stages. However, it is important to define the notation to be used along the description of each stage. The main goal in the pricing strategy is to obtain the optimal price maximizing margin. Therefore, consider a set of products $P$ of size $n$ to be sold in a convenience store. These products are grouped according to a commercial hierarchy into $m$ categories $C_i = \{p_k | k = 1, ..., n\}, \forall i = 1, ..., m$ and $p_k \in P$. Therefore, the union of all categories is the set of all products $\bigcup_{i=1}^{m} C_i = P$. Some categories will contain a large number of products e.g. a category of candies, and others will have a small group of products e.g. dairy products. Then, the category $C_i$ has a size $q_i = |C_i|$ and depends on the type of category itself.

In this case of study each product $p_k$ is clustered in a category $C_i$. Also products have historical weekly data of size $w$. The data provides information related with the date of sale such as week ($x_0^k$), year ($x_1^k$), month ($x_2^k, ..., x_{12}^k$). There is also information about weather ($x_{13}^k$), volume sales ($x_{14}^k$), discount sale ($x_{15}^k$), net cost ($x_{16}^k$), competitor price ($x_{17}^k$), and prices ($x_{18}^k, ..., x_{18+(q_i-1)}^k$). The prices of the historical data includes the price of the product $p_k$ as well as the prices of products belonging to the same category.

To be able to optimize prices, it is helpful to determine how prices influence the volume of sales and the margin. It is also important to determine how other information affects the business performance measures just mentioned in order to include them in any model. For that purpose, a preliminary analysis was performed for all categories at the same time. Results show that product sales change according to the seasons. In figure 1, we present evolution of unit sales each week of the year, for years 2013, 2014 and 2015. Maximum, mid and low sales tendencies are clear.

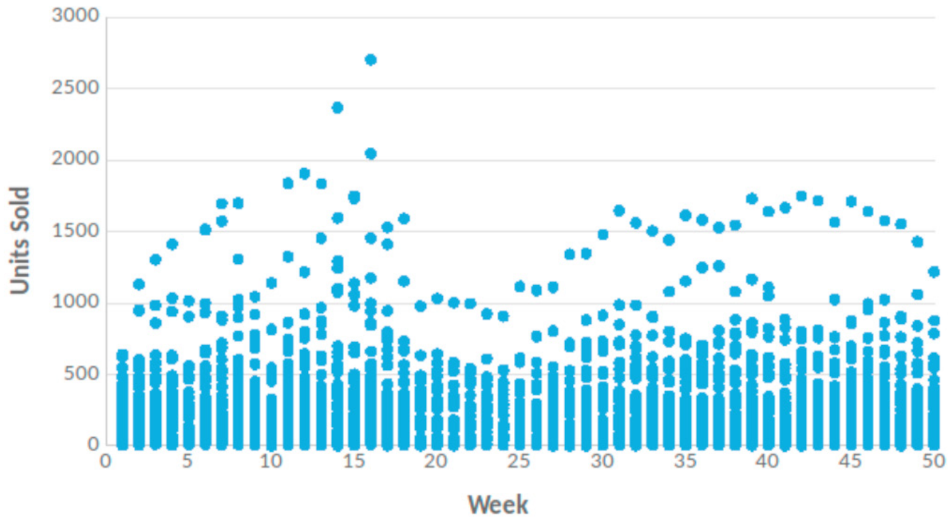


Figure 1: Seasonality of products sales of someone category

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. PCA, by finding the eigenvalues and the eigenvector of the correlation matrix, determines orthogonal axis of maximum variance. Variables are projected into those axes calling these factor loadings. Whereas observations, when projected into the axis are called factor scores. Factor loadings show how variables group together. The dimension of a dataset can be reduced when factor loadings of different variables are rather close, indicating multicollinearity (to be described later). Factor scores show how observations group together. Clusters of observations may eventually be observed suggesting similitudes and differences between groups. Details of the procedure can be found in [10] and [8].

PCA was carried out in order to determine which variables influence the margin variable the most. The results given in table 1 shows that competitors price, cost net and prices are associated with the first principal component. The year and the week are associated with the second component. The discount event, is associated with the third component. Finally, volume sales are mildly related to the category in component 4. Values in bold correspond for each variable to the factor for which the squared cosine is the largest.

Table 1: Square cosine table.

| Variable | Comp1 | Comp2 | Comp3 | Comp4 | Comp5 |
|---|---|---|---|---|---|
| **category** | 0.009 | 0.000 | 0.280 | **0.544** | 0.021 |
| **year** | 0.002 | **0.711** | 0.000 | 0.002 | 0.146 |
| **week** | 0.000 | **0.832** | 0.000 | 0.000 | 0.003 |
| **competitors price** | **0.519** | 0.001 | 0.000 | 0.030 | 0.000 |
| **net cost** | **0.906** | 0.001 | 0.023 | 0.016 | 0.001 |
| **volume sales** | 0.153 | 0.002 | 0.161 | **0.397** | 0.003 |
| **price** | **0.847** | 0.001 | 0.018 | 0.010 | 0.001 |
| **weather** | 0.000 | 0.179 | 0.000 | 0.026 | **0.773** |
| **discount** | 0.033 | 0.001 | **0.689** | 0.016 | 0.006 |

Then, we can conclude from the preliminary analysis that the features given in the historical data are relevant for the study except weather conditions. However, this last feature may result significant in the analysis by product during the first stage of the methodology (clustering analysis).

The general approach is to select the smallest subset that fulfills certain statistical criteria. In the clustering analysis, the objective is the selection of significant variables by product. The reason to use a subset of variables instead of a complete set is because the subset model might actually estimate the regression coefficients and predict future responses with smaller variance than the full model using all predictors. PCA and a Best Subset Regression (BSR) are proposed for such selection.

BSR is an exploratory model building regression analysis. It compares all possible models that can be created based upon an identified set of predictors. The general approach is to select the smallest subset that fulfills certain statistical criteria. Details of the procedure can be found in [11].

In order to select variables using PCA, we use the projection of variables to the principal component vector called $FactorLoadings$. In this way, the variables associated with the component that explain at least 80% of the variability of data is selected for the regression analysis. For BSR selection, the method is straightforward. Phase 1 ends with the selection of variables, leading to Phase 2, the MLR analysis.

The main objective in the MLR is to obtain econometric models for each product $p_k$. Therefore, the MLR attempts to model the relationship between a set of explanatory variables $\hat{X}^k = [x_1^k, ..., x_{18+q_i-1}^k]$ and a response variable $y^k$ by fitting a linear equation to observed data. Each vector has a size of $w$, where for every value of independent variables $\hat{X}^k$ is associated with a value of the dependent variable $y^k$.

The population regression line for the explanatory variables $x_1^k, ..., x_{18+q}^k$ is defined to be $\mu_y^k = \beta_0^k + \beta_1^k x_1^k + \beta_2^k x_2^k + ... + \beta_{18+q_i-1}^k x_{18+q_i-1}^k$. This line describes how the mean response $y^k$ changes with the explanatory variables. The observed values for $y^k$ vary about their means and are assumed to have the same standard deviation . The fitted values $b_0^k, b_1^k, ..., b_{18+q_i-1}^k$ estimate the parameters $\beta_0^k, \beta_1^k, ..., \beta_{18+q_i-1}^k$ of the population regression line. With all the explanatory variables, formally, the model for multiple linear regression, given $w$ observations of product $p_k$ within category $C_i$, is given by (1).

$$y^k = \beta_0^k + \sum_{s=1}^{18+q_i-1} \beta_s^k x_s^k; \qquad \forall k \in P, \qquad q_i = |C_i| \tag{1}$$

Equation (1) showed the model for the complete set of explanatory variables. However, after phase 1, not all explanatory variables are used for the regression analysis. Consequently, after the selection of the explanatory variables, formally, the model for MLR, given $w$ observations, is given by (2).

$$y^k = \beta_0^k + \sum_{s\in\{1,\dots,18+q_i-1\}} \beta_s^k x_s^k; \qquad \forall k \in P \qquad q = |C_i| \tag{2}$$

Due to the similarity of the data, multicollinearity issues may be present. In statistics, multicollinearity is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Consequently it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

In order to detect multicollinearity issues in the econometric model, the variance inflated factor (VIF) [9] is computed for each coefficient in the model. There are many different ways of addressing the issue of multicollinearity among the predictor variables, depending on its source. Ridge regression happens to be one of those methods that addresses the issue of multicollinearity by shrinking (in some cases, shrinking it close to or equal to zero, for large values of the tuning parameter) the coefficient estimates of the highly correlated variables.

Unlike least squares method, ridge regression [3] produces a set of coefficient estimates for different values of the tuning parameter. So, it's advisable to use the results of ridge regression (the set of coefficient estimates) with a model selection technique (such as, PCA or BSR) to determine the most appropriate model for the given data.

The metric to determine if a model is having high collinearity issues is a thumb rule of $VIF \geq 30$ [13].

If Ridge regression is not enough to diminish the VIF, then an iterative process to continue reduce the dimensionality of the model is used. Figure 2 shows the details of the procedure.

The third and final phase is the optimization model. The optimization phase consider the econometric models obtained during the regression analysis. In order to enrich the proposal with features of the real operation of the convenience stores, a set of business rules are included. Business rules includes operational considerations of the business as well as the external factors of the market. The optimization model is performed per category $C_i$. The mathematical formulation of the model is given in the following equations.

The objective function is given in (3), which consist on maximizes the total gross margin $y^k$ per category $C_i$. The margin is given by the econometric models obtained during the regression analysis.

$$Maximize \sum_{k\in C_i} y^k = \beta_s^k + \sum_{s\in\{1,..,17\}} \beta_s^k x_s + \sum_{s\in\{18,..,18+q\}} \beta_s x_s; \tag{3}$$

For the optimization model, the unknown variables are the total gross margin $y^k$ and the prices of the product in the category $x_s | s \in \{18, .., 18+q\}$. The known parameters are the

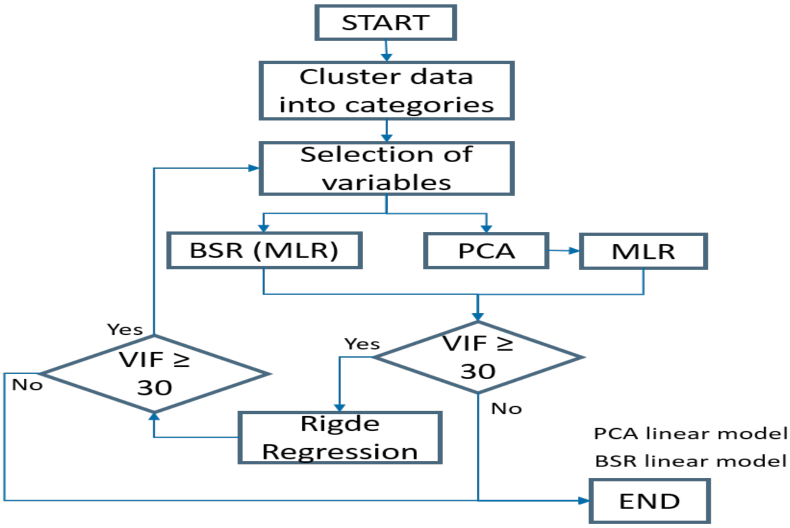


Figure 2: Iterative procedure for multicollinearity issues

coefficients computed in the regression $\beta_s^k | s \in \{1,..,18+q\}$ and the explanatory variables $x_s | s \in \{1,..,17\}$ associated with tendency, weather conditions, discounts, cost net, and seasonality. Uncertainty of the data is introduced for the parameter $x_1^k 4$ associated with volume sales. Then, $x_{14}^k \sim U(a,b)$ as it is assumed that volume sales follows a Uniform distribution, where $a,b$ are parameters desired as minimum and maximum values of volume sales according to history and experience in the business.

Additional parameters are considered in order to include business rules as constraints in the optimization. The primary business rule for a price is to overcome the net cost of the product for at least 3% (4).

$$x_k \geq 1.03 * x_{16}^k; \qquad \forall k \in C_i \in P; \tag{4}$$

Upper and lower bound are given in (5) and (6). In these equations, the proposed price should be $\pm 15\%$ of the selected reference. The options are the competitor price $x_{17}^k$ or the reference price $ref^k$. The reference price is the last price used in the convenience store. Typically, convenience stores are free to set the price of the products as appropriate to the business, however, there are products with a price controlled by the government. Also, some suppliers label the wrap of the product with a special price. In these cases, the reference price and the competitors price are set to the controlled price.

$$x_k \geq 0.85 * \min\left(x_{17}^k, ref^k\right); \qquad \forall k \in C_i \in P; \tag{5}$$

$$x_k \leq 1.15 * \max\left(x_{17}^k, ref^k\right); \qquad \forall k \in C_i \in P; \tag{6}$$

When a category of products is under analysis, it is common to find products of the same brand and same content, only differing in flavor. Therefore, those products should have the same price, unless a marketing strategy suggest differentiate one of the products. Then, we define a set of features $F_i = \{p_k | k \in 1,...n;\}$ where $F_i \in C_i$. This set contains all products under the same features and marketing strategy. An they are constraint to have the same price (7).

$$x_k = x_s; \qquad \forall k \neq s \qquad k, s \in F_i \in C_i \in P; \tag{7}$$

Another characteristic to consider for a product is the different sizes available at the store. Then, (8) considers that the price of the product $k$ should be greater than the price of the product $s$ if both are the same product under different size, where $s$ is smaller than $k$. Such definition is given in the set $S_i = \{p_k | k \in 1, ...n;\}$ where $S_i \in C_i$. This constraint also serve to differentiate equal size products but with different marketing strategy.

$$x_k \geq x_s; \qquad \forall k \neq s \qquad k, s \in F_i \in C_i \in P; \tag{8}$$

Finally, non-negativity of the variables is given in (9). It is important to highlight that prices in a convenience store are usually rounded to cents. For this case study, the prices should be rounded to $0, 0.1¢, 0.5¢, 0.9¢$, due to is the price to be label at the store.

$$x_k \geq 0 \qquad \forall k \in C_i \in P; \qquad x_k \in \{a.b\,|a \in \mathbb{N}, b \in \{0, 1, 5, 9\}\} \tag{9}$$

# 3 Solution methods

To test the performance of the proposed models, instances from [5] are tested with the proposed methodology. The instances are taken from a real case study. The data for each instance correspond to a real life case from stores located in three cities. Table 2 shows the nomenclature used to identify instances. Due to data confidentiality, the nomenclature will not show compromised information of the actual places. Then, cities are identified as $CT$. Products are classified in different levels regarding its purpose. In this research, two levels of classification are used. The first level is named "Segment" and give information about the market segment of the product. The next level is the "Category", which grouped the products by specific similarities. Categories were built during phase 1 of the proposed methodology according to the commercial hierarchy.

Table 2: Instances used for the test of the proposed methodology

| **Cities** | **Stores** | **Segment** | **Category** | **Products** |
|---|---|---|---|---|
| CT1 | 132 | S | 26 | 527 |
| CT2 | 121 | G | 46 | 407 |
| CT3 | 135 | S,G | 71 | 1090 |

A weekly historical data of the total stores from 2013 to 2015 was used for the regression analysis. Given the amount of data, a procedure to select, clean and organized data was implemented using Visual Fox Pro. The selection of explanatory variables as well as the regression analysis was implemented using a statistical software R. The linear optimization model was implemented in FICO Xpress 7.0.1, the optimization model runs until optimality.

As shown in the table 2, and according to the objectives of the company, the variety of size is good enough to prove whether the proposal is efficient for a business plan.

# 4 Results and discussion

The results of this research are presented in terms of reduction of the dimensionality and the quality of the econometric model versus the approach from [5]. Then, table 3 shows the reduction of dimensionality of the econometric models obtained. Rows show the minimum, average, maximum percentage of reduction obtained in the econometrics models according to the applied procedure. As can be noted, the reduction from BSR procedure is the lower among the others.

Table 3: Percentage of reduction of dimensionality

| Metric | CITY | Hervert, 2016 | PCA | BSR |
|---|---|---|---|---|
| **Min** | CT-1 | 44% | 89% | 88% |
| | CT-2 | 61% | 67% | 47% |
| | CT-3 | 53% | 60% | 44% |
| **Average** | CT-1 | 82% | 90% | 89% |
| | CT-2 | 89% | 88% | 73% |
| | CT-3 | 91% | 84% | 77% |
| **Max** | CT-1 | 90% | 95% | 91% |
| | CT-2 | 89% | 92% | 67% |
| | CT-3 | 99% | 95% | 91% |
| **Mode** | CT-1 | 87% | 89% | 89% |
| | CT-2 | 98% | 96% | 85% |
| | CT-3 | 93% | 83% | 89% |

Quality of the econometric models can be measure through the variance inflated factor. The result is given in Table 4, where the rows show the minimum, average and maximum VIF found in the econometric models. These result is given according to the applied procedure.

Table 4: Variance Inflated Factor

| Metric | CITY | Hervert, 2016 | PCA | BSR |
|---|---|---|---|---|
| Min | CT-1 | 13.85 | 1.00 | 1.47 |
| | CT-2 | 14.87 | 1.00 | 1.30 |
| | CT-3 | 12.42 | 1.00 | 1.27 |
| Average | CT-1 | 325.42 | 3.03 | 9.07 |
| | CT-2 | 184.23 | 3.28 | 8.55 |
| | CT-3 | 356.42 | 4.03 | 11.58 |
| Max | CT-1 | 10268.05 | 21.96 | 29.95 |
| | CT-2 | 5430.37 | 29.33 | 534.24 |
| | CT-3 | 7383.23 | 26.03 | 1565.88 |

The PCA procedure obtained the best result of VIF. Which indicates that PCA allows to select the best subset of variables without carrying multicollinearity issues. Both, PCA and BSR performed better than the approach from [5].

Along with the VIF, the R squared in another metric that allows to measure the quality of the econometric models. R-squared is a handy, seemingly intuitive measure of how well your linear model fits a set of observations. The result is given in Table 5, where the rows show the minimum, average and maximum R squared found in the econometric models. These result is given according to the applied procedure.

Table 5: R-Squared result

| Metric | CITY | Hervert, 2016 | PCA | BSR |
|---|---|---|---|---|
| Min | CT-1 | 0.01 | 0.08 | 0.45 |
| | CT-2 | 0.02 | 0.00 | 0.51 |
| | CT-3 | 0.01 | 0.01 | 0.39 |
| Average | CT-1 | 0.66 | 0.79 | 0.97 |
| | CT-2 | 0.73 | 0.61 | 0.96 |
| | CT-3 | 0.67 | 0.75 | 0.96 |
| Max | CT-1 | 0.97 | 0.99 | 0.99 |
| | CT-2 | 0.98 | 1.00 | 1.00 |
| | CT-3 | 0.97 | 0.98 | 0.99 |

For the R-squared, the BSR procedure obtained the best result. R-squared provides an estimate of the strength of the relationship between the model and the response variable. Both, PCA and BSR performed better than the approach from [5].

As for the optimization model, due to confidentiality of data, we are not able to present the entire result. However, in figure 3 we show an example for a category. The results are given in terms of price and margin per product in a category. The confidentiality agreement of the case of study prevent us to show any number related, then, scales were removed from the chart. However, the purpose of the chart is to show the effect of the analysis compared with current methods.

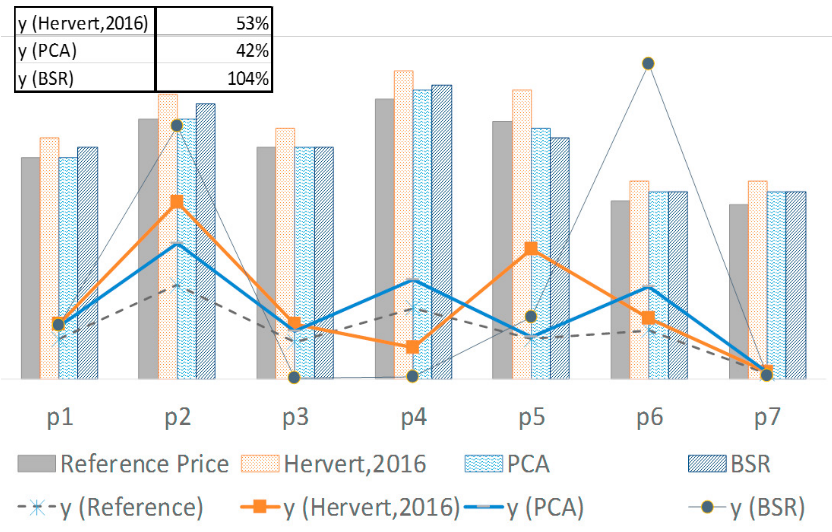


Figure 3: Example of the result of optimization for one category. *note: Scales were removed from the chart due to confidentiality issues.*

In this figure, bars represent the price proposed by each procedure, and it is compared to a reference. The reference correspond to the real price used in the store, whereas the others are the optimal prices that should be used instead. Lines correspond to the margin result per product, it is clear that halo and cannibalization effects are string in the BSR models. Additionally, the variable for volume sales plays and important role for the proposal. During the implementation, the strategy was to set the values for the Uniform distribution of the volume sales as a desire parameter of minimum and maximum volume to achieve for the same period.

# 5 Conclusions

Convenience stores serve the entire purchasing population of its geographical area but focuses on customers who need to purchase items outside of normal working hours such as swing shift employees and quick shoppers looking for snacks and related items. In order to capture attention and sales they use prominent signs at the store locations, billboards, media bites on local news, and radio advertisements to capture customers. These actions combined with a right price strategy will lead to the success of the business.

In this research we proposed a pricing strategy composed for three phases. The objective in the first phase was to understand and use the most appropriate variables for the business. For this purpose a PCA and BSR procedures are used. In order to solve multicollinearity issues, an iterative reduction procedure with Ridge regression was implemented. Later, through a MLR we were able to obtained econometric model that will better focus in the major goal, the increase of the margin. Then, the optimization model will take additional business rules (internal and external) in order to obtain a price proposal that allowed to increase the margin under the most possible real scenario. In this way, uncertainty of the volume sales was added.

The proposed methodology is easy to understand and implement within most fields of business. Also, it produces information that can be used to create forecasting and planning models and it is relatively inexpensive research process if a firm has high quality data. Also, the use of business rules give an advantage over typical econometric models that only considers one price and not crossed information.

Future work in this research includes the possible application of new methodologies in a pilot test. These pilot will help to featuring better the uncertainty data from some elements of the process.

# Acknowledgments

The authors are grateful to Sintec for financial and technical support during the development of this research. Sintec is the leading business consulting firm for Supply Chain, Customer and Operations Strategies with a consultative model in Developing Organizational Skills that enable their customers to generate unique capabilities based on processes, organization and IT. Also, we appreciate the financial support of CONACYT-SNI program in order to promote quality research.

# References

[1] M. Abuella and B. Chowdhury. Solar power probabilistic forecasting by using multiple linear regression analysis. In *SoutheastCon 2015*, pages 1–5, April 2015.

[2] Shabri Ani and Samsudin Ruhaidah. Crude oil price forecasting based on hybridizing wavelet multiple linear regression model, particle swarm optimization techniques, and principal component analysis. *The Scientific World Journal*, 2014, 2014.

[3] M. El-Dereny and N. I. Rashwan. Solving multicollinearity problem using ridge regression models. *International Journal of Contemporary Mathematical Sciences*, 6(12):585–600, 2012.

[4] Chun H. and Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society*, 72(1):3–25, 2010.

[5] Lopez-Perez Jess Fabin Hervert-Escobar Laura and Esquivel-Flores Oscar A. Optimal pricing model: case of study for convenience stores. *Lecture Notes in Artificial Intelligence*, page 8, 2016.

[6] WebFinance Inc. pricing strategy, 2016.

[7] Z. Ismail, A. Yahya, and A. Shabri. Forecasting gold prices using multiple linear regression method. *American Journal of Applied Sciences*, 6:1509–1514, 2009.

[8] Ian T. Jolliffe. *Principal Component Analysis (second edition)*. Springer, 2002.

[9] Dan Liao and Richard Valliant. Variance inflation factors in the analysis of complex survey data. *Survey Methodology*, 38(1):53–62, 2012.

[10] Y. Ma, S. Sastry, and R. Vidal. *Generalized Principal Component Analysis*. Interdisciplinary Applied Mathematics. Springer New York, 2015.

[11] Alan J. Miller. Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, 147(3):389–425, 1984.

[12] Avval Zhila Mohajeri, Pourbashir Eslam, Ganjali Mohammad Reza, and Norouzi Parviz. Application of genetic algorithm - multiple linear regressions to predict the activity of rsk inhibitors. *Journal of the Serbian Chemical Society*, 80:187–196, 2015.

[13] Robert M. O'brien. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690, 2007.

[14] Xiaozhen Yan, Xie Hong, and Wang Tong. A multiple linear regression data predicting method using correlation analysis for wireless sensor networks. In *Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), 2011*, volume 2, pages 960–963, July 2011.

[15] Li-Ping Zhu, Xiangrong Yin, and Li-Xing Zhu. Dimension reduction via an alternating inverse regression. *Journal of Computational and Graphical Statistics*, 19(4):887–899, 2010.

[16] Li-Ping Zhu, Li-Xing Zhu, and Zheng-Hui Feng. Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466, 2010.