

Data Mining and Knowledge Discovery Approach for Manufacturing in Process Data Optimization

Raed S. Batbooti and R.S. Ransing

Abstract The in-process data offers a rich source for discovering new process knowledge. This is particularly important in manufacturing processes such as casting process where the number of process variables are large ($\sim 30\text{--}50$) and process observations are small ($\sim 50\text{--}60$). A multivariate data analysis technique such as principal component analysis (PCA) is used for discovering knowledge from foundry in-process data. The correlations among factors for a given response are discovered by projecting the data on a reduced dimensional space defined by the principal components. The correlations are discovered among both categorical and continuous data. A new methodology has been introduced that uses scores and loadings in the PCA to define optimal and avoid tolerance limits for factors. Interactions among factors are also considered. The developed approach allows process engineers to adjust system parameters as it discovers factor tolerance limits that contribute most to the overall variance. This information is used to suggest corresponding optimal and avoid limits that would result in the reduction of variance. The workings of the algorithm are demonstrated on a foundry case study.

1 Introduction

PCA is a multivariate technique widely used in different applications to reduce dimension, compress data, simplify data description and extract the most significant information by projecting the data from a higher dimensional space into a sub linear space [1, 2]. The resulted sub space can be used as a dimension reduction space and

R.S. Batbooti (✉) · R.S. Ransing
Swansea University, Swansea SA2 8PP, UK
e-mail: 591619@swansea.ac.uk

R.S. Ransing
e-mail: r.s.ransing@swansea.ac.uk

R.S. Batbooti
University of Basra, Basra, Iraq

analysis space to infer the relation between variables. This way the PCA transfers the problem from a high dimensional manifold into a linear combination of original parameters known as principal components. This analysis is based on maximum data variance and minimum projection cost.

For manufacturing processes, production goals such as maximization of mechanical properties and/or minimization of mechanical defects are often defined. Such analysis requires using process knowledge to change factor settings in order to optimize the process. This work builds on a new co-linearity index approach based on PCA and penalty matrix approach to reduce production defects [3]. The co-linearity indices are derived from loadings of PCA and used to infer a potential correlation between a response and the associated factors. The work is extended to include categorical variables such as week days by using multiple factor analysis (MFA) as pre-treatment tool [4]. The poster will present the formulation and demonstrate the further development of algorithms to use scores and loadings for finding optimal ranges of factors that are likely to minimize occurrence of defects.

2 Method

In PCA variables represented by loadings and observations are represented by scores in the resulted sub space. However, instead of using loading and score plots, a co-linearity plot and projected scores plots is used to find out the correlation between variables and the contribution of observations on variables respectively. The in-process data consist of m observations and n variables which include one response at least and the remaining is factors. In the case study example shown in the poster, the main aim of the analysis is to reduce the incidence of conchoidal fractured surface area in a steel casting. The fracture surface represents the response, which characterized by 19 continuous variables represent the factors.

2.1 Co-linearity Index (CLI)

The co-linearity index (CLI) is a visual tool to find the relation between specific responses with factors to understand which factors effects a specific response, which is very beneficial to process engineers based on p principal components instead of 2 or 3 principal components in convensional PCA. It can be calculated by using the following steps [3]:

1. Data pre-treatment: centering and standardization are used to preserve the variance and scale the data respectively. Another three sets of transformation that are important in this approach are penalty value transformation for responses (where the response values are transferred to the range $[0 \ 1]$ according to maximum and minimum penalty value thresholds chosen by the analyst),

Multiple Factor Analysis (MFA) to transfer categorical variables into continuous variables (by using an indicator matrix in it each variable is replaced by a set of indicator binary variables taking one if the categorical variable has been observed and zero otherwise) and median-interquartile range transformation for quantitative variables. All above transformation followed by centring and scaling the data by extracting the mean value and divide each column on its Eigen value respectively.

2. Apply PCA on covariance matrix resulted before step (1). $Cov = \frac{1}{n-1} X_T^t$.
 3. Estimate the loading matrix based on the following equation $L_s = D_s^{-1} V D_e$ Where: L_s is the standardized loading matrix, V is the matrix of eigenvectors arranged as column vectors in descending order of eigenvalues, D_s is the diagonal matrix of the standard deviations of the columns of X_T and D_e is the diagonal matrix containing the square roots of eigenvalues.
 4. Evaluate the correlation matrix from $L_s L_s^t$ for p principal components, where the inner product of i th and j th row vectors of L_s represents the correlation between variable i and j . After that co-linearity index can be plotted by plotting angles and length of the loading vectors. Scree plot method is used to choose the optimal number of principal components.
 5. Divide the co-linearity index plot into five regions
 - The no correlation region between -0.2 to 0.2 co-linearity index.
 - The two weak correlation regions between -0.5 to -0.2 and 0.2 – 0.5 respectively.
 - The two strong correlation regions, which include co-linearity index between -1 to -0.5 and 0.5 – 1 for negative and positive correlation respectively.
- The CLI plot for steel alloy is displayed in Fig. 1. Where there are five factors correlated with high penalty direction (Pouring temperature (F), %Cr, %P, %S and %Ti) and three factors showed a correlation with low penalty direction (%Zr, Mn/S Ratio and Carbon Drop).

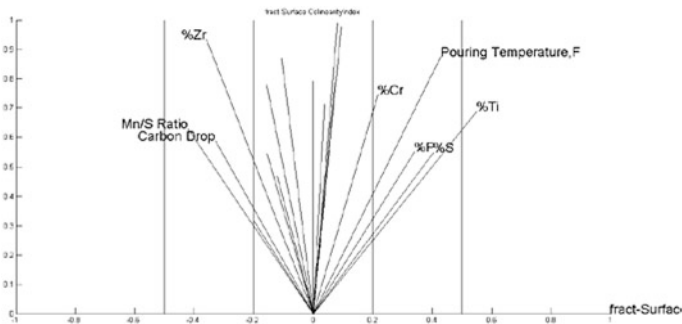


Fig. 1 Co-linearity index plot for the in-process data used in steel alloy

2.2 Scores Projected Space to Predict the Optimal Variables Range and Variables Recommendation

The main steps for calculating optimal system settings for quantitative and categorical variables are summarised below [5]:

1. Find the correlated variables from applying CLI as described in Sect. (2.1).
2. Create a contribution plot for each variable. The horizontal axis of the plot represents the projected scores on the variable, whereas the vertical axis represents the projected scores on the response. The scores lay on the resulted new subspace bounded by positive variable projection axis and corresponding response direction it represents. It should be noted that the response direction is the direction of correlation with the variable. From linear algebra, the projection of score t_i on loading L_j in p dimensions is expressed as:

$$t_i^* = \frac{\sum_{k=1}^p L_j(k) * t_i(k)}{\sqrt{\sum_{k=1}^p (L_j(k))^2}} = \frac{L_j \cdot t_i}{\|L_j\|}$$

3. These scores relate to either optimal or avoid ranges with reference to the correlated variable. The observations corresponding to the collected scores is stored in either variable $x_{optimal}^j$ or x_{avoid}^j depending upon whether the correlation is positive or negative and the number of observations stored are counted and stored in a variable n_x^j .
4. Determine the range for factors using the observations stored in $x_{optimal}^j$ or x_{avoid}^j .
5. For categorical variables determine the percentage of occurrences P_o^j :

$$P_o^j = \frac{\sum_{i=1}^{n_x^j} x(optimal)_i^j}{n_x^j} \times 100 \% \text{ or } P_o^j = \frac{\sum_{i=1}^{n_x^j} x(avoid)_i^j}{n_x^j} \times 100 \%$$

where n_x^j = number of elements in x^j .

The categorical variable j is chosen for recommendation as optimal classification if $P_o^j \geq 60 \%$ and is negatively correlated with penalty values high penalty values. The recommendation will be ‘avoid’ for a categorical variable i if $P_o^j \geq 60 \%$ and the variable is positively correlated with high penalty values. Figure 2 shows the main steps of finding the optimal variables range. Score projection and optimal system settings plot displayed in Figs. 3 and 4 respectively.

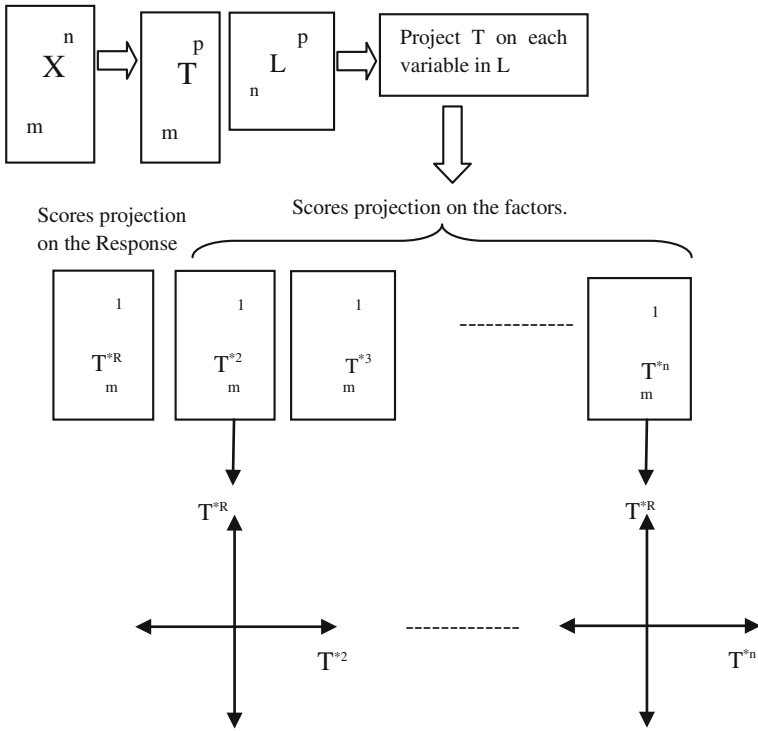


Fig. 2 Scores projection to predict the optimal process settings

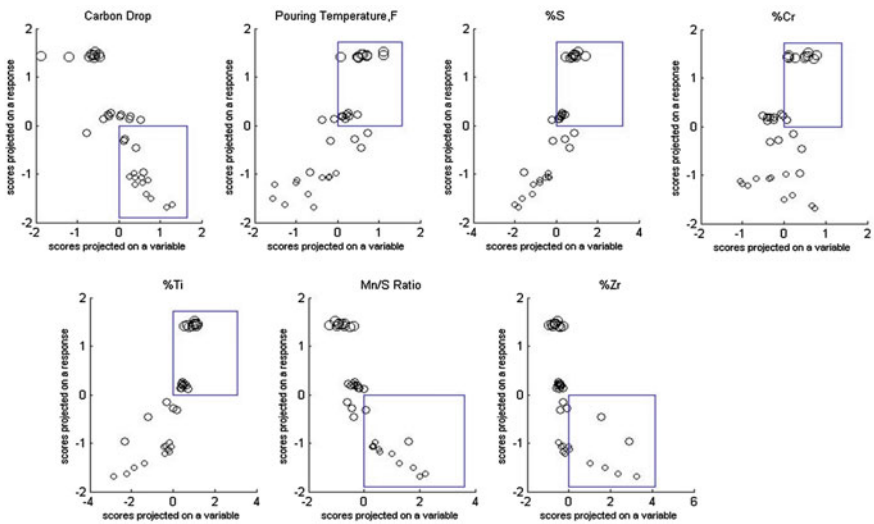


Fig. 3 Scores projection on variables and response of steel alloy

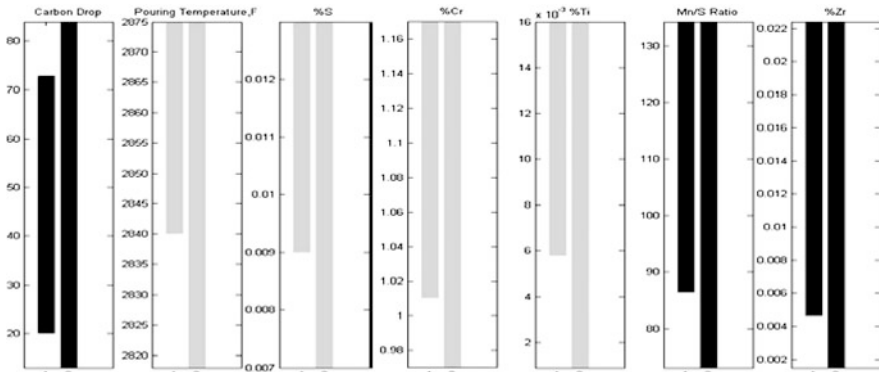


Fig. 4 The optimal range for variables, for each variable the left hand bar represent the optimal (black bars) or avoid (the light bar) range, obtained range for each variable corresponding to scores bounded by rectangle in Fig. 3 for the variable

3 Conclusion

In the proposed work, an enhanced co-linearity index procedure is used to predict correlations among factors for a given response using foundry in-process. A new approach has been proposed to predict the optimal process settings for correlated variables by using the analogue between loading and scores of principal component analysis as used in the bi-plots. The concept is however extended for p number of principal components. The procedure and results of the algorithm are presented in the context of a foundry case study.

References

1. Bishop, C.: Pattern Recognition and Machine Learning. Springer, New York (2006)
2. Abdi, H., Williams, L.: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, pp. 433–459 (2010)
3. Ransing, R., Giannetti, C., Ransing, M., James, M.: A coupled penalty matrix approach and principal component based co-linearity index technique to discover product specific foundry process knowledge from in-process data in order to reduce defects. *Comput. Ind.* **64**(5), 514–523 (2013)
4. Giannetti, C., Ransing, R., Ransing, M., Bould, D., Gethin, D., Sienz, J.: A novel variable selection approach based on co-linearity index to discover optimal process settings by analysing mixed data. *Comput. Ind. Eng.* **72**, 217–229 (2014)
5. Ransing, R.S., Batbooti, R.S., Giannettia, C., Ransing, M.R.: A quality correlation algorithm to minimise unexplained deviation from expected results in production batches. *Comput. Ind. Eng.* (2015) (submitted)