ISIJ International, Vol. 45 (2005), No. 12, pp. 1943–1945

Note

Prediction of Silicon Content in Blast Furnace Hot Metal Using Partial Least Squares (PLS)

Tathagata BHATTACHARYA

Research & Development Division, TATA Steel, Jamshedpur, Jharkhand, PIN 831001, India.
E-mail: tathagata.bhattacharya@tatasteel.com

(*Received on July 11, 2005; accepted on September 13, 2005*)

1. Introduction

Multivariate modeling has gained popularity in several process industries, especially in the petrochemical sector. The Partial Least Square (PLS) is one of the various multivariate techniques where the relationship between multiple Y (responses) and large number of X variables (predictors) are modelled. Recently, the technique has been used in the steel industry.[1–4] The PLS is emerging as the most robust[4,5] and reliable prediction tool when huge amounts of collinear data are to be handled. Collinear data means any two columns in the data set are linearly dependent and the inverse of the matrix is non-existent since the determinant is zero. The other multivariate techniques like multiple linear regression (MLR) fails to handle collinear data as it involves inversion of matrix to estimate the regression coefficients.

The present work deals with the application of PLS to predict the silicon content of hot metal. This is a novel application of PLS in ironmaking since PLS has traditionally been used in many other scientific fields like chemometrics, chemistry, biology *etc.* to name a few.[5,6] Use of PLS in hot metal silicon prediction is a novel application since most of the reported Si prediction models have employed ANN[7–10] and nonlinear time series methods.[11–13]

2. Methodology

First introduced by Wold[14] in the field of econometrics, PLS has become an important technique in many areas and especially in process control and process monitoring. It is a method for relating two data matrices, **X** and **Y**, by a linear multivariate model, but goes beyond traditional regression in that it also models the structure of **X** and **Y**.[6] The very nature of the PLS algorithm (outer relation, inner relation, exchange of scores, mixed relation *etc.*) itself guarantees a good predictive model.[5,15] This will be discussed in Sec. 2.3.

2.1. Pre-processing of Data

The analysis, like any other data modeling technique, consists of two steps-training and prediction steps. Before the model is developed, it is convenient to tailor the data in the training step to make the calculations easier by ensuring zero mean and unit standard deviation for each column. In the present analysis, a standard Mahalanobis scaling has been employed prior to model building. For example, any data value x_i has been replaced by its scaled value x_i^{scaled} according to the relation: $x_i^{\text{scaled}}=(x_i-\mu(x))/\sigma(x)$, where $\mu(x)$ and $\sigma(x)$ are mean and standard deviation of all data for the variable x.

2.2. Principal Component Analysis (PCA)

The PLS is basically built on the concept of principal component analysis (PCA). So, a good understanding of PCA is helpful in interpreting the results from PLS. PCA is a method of writing a matrix **X** of rank r as a sum of r matrices of rank 1 as below:

$$\mathbf{X}=\mathbf{M}_1+\mathbf{M}_2+\cdots+\mathbf{M}_r \quad \text{..... (1)}$$

Rank is a number expressing the true underlying dimensionality of a matrix. A collinear matrix will have a rank less than the number of columns present in the matrix. These rank 1 matrices, $\mathbf{M}_h$, can all be written as outer product of two vectors, a score $\mathbf{t}_h$ and a loading $\mathbf{p}_h'$:

$$\mathbf{X}=\mathbf{t}_1\mathbf{p}_1'+\mathbf{t}_2\mathbf{p}_2'+\cdots+\mathbf{t}_a\mathbf{p}_a'=\sum \mathbf{t}_h\mathbf{p}_h' \quad \text{.....(2)}$$

or the matrix equivalent $\mathbf{X}=\mathbf{TP}'$. Here we can deflate the matrix **X** up to a number of component matrices where $a\leq r$. The scores and loadings are the projections of rows and columns of the data matrix onto a single dimension respectively. For further physical interpretations of the scores and loadings, the reader may refer to Geladi *et al.*[5] Amongst other algorithms,[16] the nonlinear iterative partial least squares (NIPALS)[5] algorithm is used to calculate the component matrices. The scores are nothing but the eigenvectors of the matrix **X**. Therefore, NIPALS basically computes the eigenvectors. The component matrices are arranged so that the first one (*i.e.*, $\mathbf{M}_1$) contributes mostly to the variation of the data matrix, the second component matrix has the second largest contribution and so on.

2.3. Partial Least Squares (PLS)

In PLS model building, mainly three major steps[5] are followed to calculate the model parameters. The three relations are called outer relation, inner relation and mixed relation. The outer relation is nothing but the decomposition or deflation of **X** and **Y** data matrices into their principal components *i.e.*, obtaining scores and loadings for **X** and **Y** data matrices:

$$\mathbf{X}=\mathbf{TP}'+\mathbf{E}=\sum \mathbf{t}_h\mathbf{p}_h'+\mathbf{E} \quad \text{.....(3)}$$

$$\mathbf{Y}=\mathbf{UQ}'+\mathbf{F}=\sum \mathbf{u}_h\mathbf{q}_h'+\mathbf{F} \quad \text{.....(4)}$$

E and **F** are the residual matrices which are to be minimised. This step is identical to PCA except the fact that in PLS both **X** and **Y** matrices are deflated whereas in PCA only the **X** matrix is deflated. The principal components in the case of PLS are called Latent Variables (LVs). The latent variables are not same as original X variables, but they are connected with the original variables through the loading vector. Therefore, the general idea of PLS is to try to extract these latent variables, accounting for as much of the manifest factor variation as possible while modeling the responses well. Like in PCA, the NIPALS algorithm is also

used here. In the inner relationship, the scores of both **X** and **Y** data matrices are exchanged and linked to give rotated components for **X** and **Y** block and this is done to give better predictive power to the model. The mixed relationship is the final step in the model building. Here the model parameter or the regression coefficient obtained from the inner relationship is used for final prediction. Since no matrix inversion is required, highly collinear data can be used for analysis in PLS. The steps and the algorithm for developing a PLS model are illustrated in Wise *et al*.[15]

The number of latent variables to be extracted depends upon the prediction error. This is typically done by cross-validation.[17] Cross-validation is a procedure where the available data is parted into training (calibration) and test (prediction) sets. The prediction error (called PRESS, the prediction residual sum of squares) on the test samples is determined as a function of the number of latent variables (LVs) extracted. The optimum number is the number of LVs which produces minimum prediction error.

3. Case Study

3.1. The Data Set

For the current study, hourly data for one month has been collected from 'G' blast furnace of TATA Steel. The dataset contains 120 *X* type variables and only one type of *Y* variable *i.e.*, the % silicon content in hot metal. The *X* type variables include process variables such as coke rate, coal rate, amount and chemistry of other raw materials, hot metal and slag, furnace instrument data, various probe data, tuyere parameters and some derived parameters such as RAFT, permeability, gas index *etc*. The initial 30 days' hourly data (30×24=720) are used as the training set and the next day's data (1×24=24) are used for prediction. Before it was used in modeling, the data matrices had undergone pre-processing as discussed in Sec. 2.1.

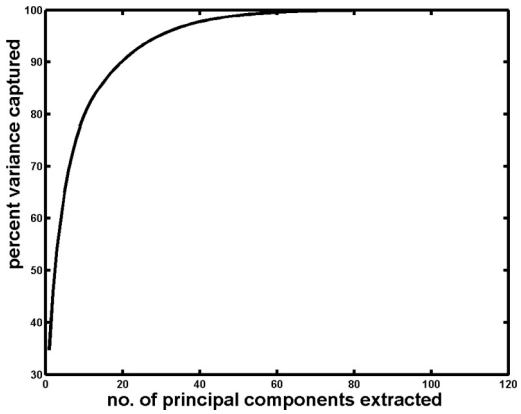

Fig. 1. Percentage variance captured by PCA model (Bhattacharya).

3.2. Modeling and Analysis

Figure 1 shows the results of PCA performed on the training data. It is found that about first 20 principal components (PCs) capture nearly 90% of the information in the data set. So, majority of the information in the data set is retained in first 20 principal components. Therefore, the 120 variables can be replaced by only 20 new variables, which are again linear combinations of the original variables (refer to loadings in PCA), with little loss of information from the original **X** data set. One interesting finding is that the last 100 small components carry only 10% information which can be attributed to noise in the original data. The loadings plot for the first PC (or the first eigenvector) is shown in **Fig. 2**. It is evident that a few process variables (around 21) contribute most to the first PC scores and hence, contribute to the first PC matrix *i.e.*, $\mathbf{M}_1$ [Eq. (1)]. The first PC, in turn, explains the maximum variation in the data. These process variables correspond to coke rate, central working index (CWI), upper furnace differential pressure, gas index at centre, a particular hearth thermocouple temperature, %C in HM, mid-furnace permeability, sinter basicity (B2), %MgO in slag and temperatures of all 12 above burden probes. Therefore, variation in hot metal silicon is also strongly related to these variables as the final PLS model relates the *Y* and *X* variables. The PCA, thus, helps in identification of key variables and reduction of data.

For building the PLS model, cross validation is to be done on the training set to determine the optimum number

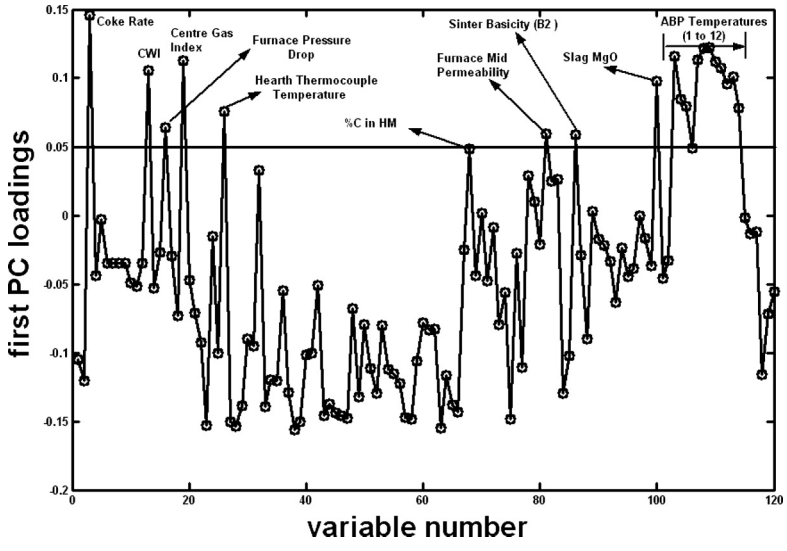

Fig. 2. Plot for first principal component loadings obtained by PCA for training data set (**X**). The horizontal line corresponds to a loading of 0.05. All the variables above this line are relatively dominant variables (Bhattacharya).

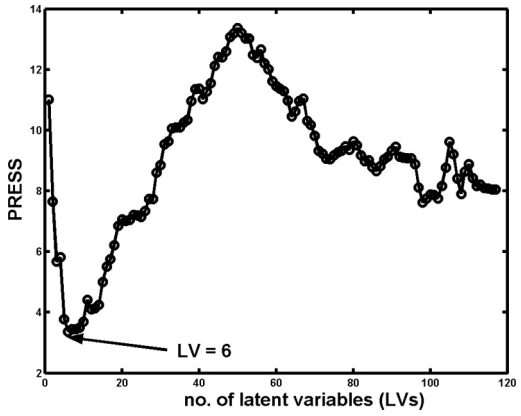

Fig. 3. The variation of PRESS against the no. of latent variables extracted in the PLS model (Bhattacharya).

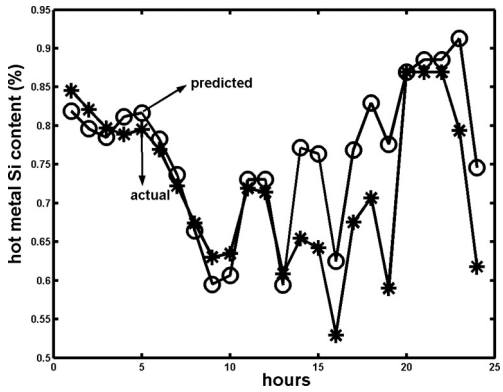

Fig. 4. The variation of actual and predicted silicon content with time (Bhattacharya).

latent variables (LVs) to retain in the model. In the present case, the PRESS (**Fig. 3**) goes through a clear minimum at 6 LVs. As more and more LVs are extracted the prediction capability of the PLS model deteriorates because of inclusion of insignificant variables and noises. **Figure 4** shows the plot of actual and predicted hot metal Si with time. The prediction has been made using 6 LVs. The hourly prediction was done for the 31st day (24 h) from the hourly training data of previous 30 d. There is a good agreement for the initial 13 h and then the predictive power of the model diminishes slightly. The trend of predicted silicon content follows the actual values all along the time span. From the hourly prediction of the model, we can find out a more reliable daily average of hot metal silicon content. For the present case, the actual day average of silicon content on the 31st day was 0.72 and the predicted day average is 0.76 (5.5% higher). Therefore, the PLS model can be of great help in monitoring the hourly prediction of silicon as well as in estimating the next day's average silicon content a priori so that the appropriate control actions could be initiated well in advance.

The predictive power could be enhanced by judiciously choosing the training and prediction interval. The process variables have to be selected correctly. Moreover, the data set has to be proper with less outlier and no missing values.

4. Conclusions

The partial least squares (PLS) technique has been successfully employed for prediction of silicon content in hot metal. The PLS exploits the structure of the data sets and helps in identifying the dominant process variables. It can handle large number of highly collinear data as well as help in data reduction.

REFERENCES

1) V. Vaculik and I. Miletic: Proc. of Int. Symp. on Control and Optimization in Minerals, Metals and Materials Processing, ed. by D. Hodouin, C. Bazin and A. Desbiens, Met. Soc., Canada, (1999), 115.
2) A. G. Taylor: IFAC Automation in Mining, Mineral and Metal Processing, ed. by J. Heidepriem, Pergamon Press, Cologne, (1998), 223.
3) S. L. Quinn and V. Vaculik: *AISE Steel Technol.*, October (2002), 37.
4) T. Bhattacharya, S. Nag and S. N. Lenka: *Tata Search*, (2004), 215.
5) P. Geladi and B. R. Kowalski: *Anal. Chim. Acta*, **185** (1986), 1.
6) S. Wold, M. Sjöström and L. Eriksson: *Chemom. Intell. Lab. Syst.*, **58** (2001), 109.
7) V. R. Radhakrishnan and A. R. Mohamed: *J. Proc. Cont.*, **10** (2000), 509.
8) M. Alaraasakka, O. Ritamaki and R. Kanniala: Ironmaking Conf. Proc., ISS-AIME, Pittsburgh, (2000), 477.
9) Z. Guangqing, M. Jitang and B. Bo: Ironmaking Conf. Proc., ISS-AIME, Pittsburgh, PA, (1996), 211.
10) Y. Bin and Y. Tianjun and Ning Xiaojun: *J. Univ. Sci. Technol. Beijing*, **7** (2000), 269.
11) H. Saxen and L. Karilainen: Ironmaking Conf. Proc., ISS-AIME, Warrendale, PA, (1992), 185.
12) M. Waller and H. Saxen: *ISIJ Int.*, **42** (2002), 316.
13) I. F. Kurunov, A. V. Ganchev, L. A. Fursova and O. V. Lagutina: *Steel USSR*, **20** (1990), 10.
14) H. Wold: Research Papers in Statistics, ed. by F. David, Wiley, New York, (1966), 411.
15) B. M. Wise and N. B. Gallaghar: *J. Proc. Cont.*, **6** (1996), 329.
16) B. S. Dayal and J. F. MacGregor: *J. Chemometrics*, **11** (1997), 73.
17) P. Geladi, and B. R. Kowalski: *Anal. Chim. Acta*, **185** (1986), 19.