

Bayesian statistical analysis of an efficiency-quality index in a chemical industrial process: a predictive analysis.

Analisi Bayesiana di un indicatore di resa-qualità in un processo chimico industriale: un'analisi predittiva

Patrizio Frederic¹

Dipartimento di Scienze Statistiche
Università di Bologna

Mauro Catellani

Stabilimento di Pontecchio Marconi
CIBA S.C.

Riassunto: Proseguendo nella collaborazione di CIBA S.C. con il Dipartimento Statistica di Bologna è stata analizzata, da un punto di vista statistico, una fase intermedia del processo di produzione di un additivo nello stabilimento CIBA sito in Pontecchio Marconi (BO). La ricerca è stata articolata in due fasi. Una fase preliminare di analisi dei dati: si sono esplorati i legami di dipendenza tra le 70 variabili coinvolte che CIBA raccoglie elettronicamente sul proprio impianto. Una seconda fase costituita dalla costruzione di un indicatore di resa-qualità dell'impianto e dalla realizzazione di un modello logit-additivo che esprime le variazioni dell'indicatore in funzione delle variazioni di alcune delle variabili del processo.

Keywords: chemical industrial process, additive models, heteroscedasticity, bayesian modelling, Markov Chain Monte Carlo samples, efficiency-quality index.

1. Introduction

Technologists of the Bologna factory of Ciba S.C. and statisticians of the Department of Statistical Sciences of the University of Bologna have begun to collaborate on the statistical analysis of a CIBA's industrial processes. For a few years, CIBA has been conducting electronic monitoring the synthesis-distillation process of one of their intermediate products. For confidential reasons such a product will be conventionally named Z .

The aim of this paper is to study the process from a statistical point of view. The need to combine a statistical study with chemical-engineering considerations derives from the intrinsically dynamic nature of the process. More than 70 variables are involved in the process and the features of recycling and refluxing in the process increase the complexity of the system. Consequently standard deterministic based model (see for example Clarke *et al.* (1987), Garcia *et al.* (1989), Bequette (1991), Coughanowr (1991)) fail in describing the process. In this paper some statistical investigations of the process are proposed to search for relations among the variables. Then a predictive distribution of an index of quality and efficiency of the final output conditioned to some variables that affect the process (temperatures, pressures, etc.) is formulated.

2. The Process

Roughly speaking the production of Z is divided in two step (a more detailed introduction to chemical industrial process can be find in Kirk (1998)). The first step, named synthesis,

¹Indirizzo per corrispondenza: frederic@stat.unibo.it

consists in mixing two components say C_1 and C_2 and subject them, for a given number of hours, to controlled combination of temperature and pressure. Under these conditions the initial mixture of C_1 and C_2 *changes* into a new one consisting of the product Z and some residuals W . The second step consists in distilling this new mixture in order to separate component Z from residuals W . In the distillation process the mixture is subjected to changes of pressure and temperature along a path consisting of three distillation columns. As a consequence of these changes, the heavier constituents tend to move toward the bottom of each column and the lighter ones toward the top. The excessively heavy (or light) constituents are eliminated as waste. Part of the remaining mixture becomes the input of the subsequent column, and the rest is recycled in the current column.

The data used in this study are divided into two groups: 68 process variables such as flows, temperatures, pressures and densities of mixture components in the column, and final analysis which are measurement of the final product output. The process variables were automatically recorded every 10 minutes from 00:00:00 hours on 1 Feb. 2002 until 23:50:00 hours on 1 Aug. 2002 (128 days). The final analysis were recorded less regularly.

The quality Q of final output Y is the percentage of product Z in Y . The efficiency E of the process is the ratio between the actual quantity of Z in Y and the theoretical quantity of Z one would obtain if the process was perfect. Whereas variable Q is one of the final analysis, variable E has to be calculated.

3. Explorative Data Analysis

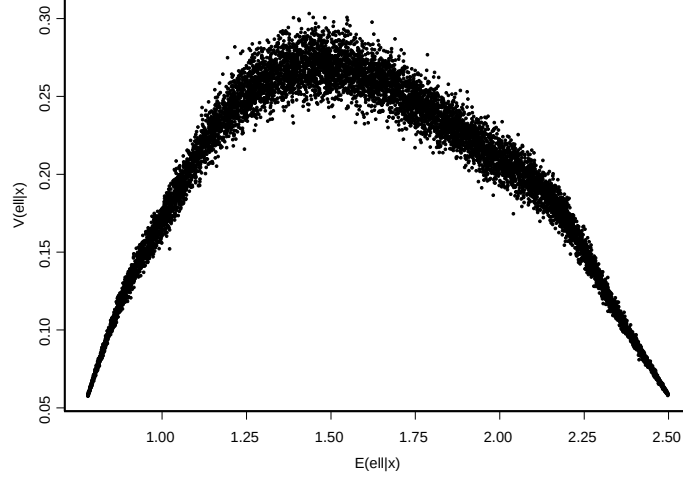
Let p be a contiguous period of time, and let $K_l(p)$ be the total number of kilograms of component C_l , $l = 1, 2$. Let $Z^t(p) = g(K_1(p), K_2(p))$ be the theoretical total number of kilograms of Z if efficiency was maximum. Let $Z^a(p)$ be the actual total number of kilograms of Z produced in period p . The efficiency for period p , say $E(p)$ is then: $E(p) = Z^a(p)/Z^t(p)$. Due to measurement error problems effective calculation of $E(p)$ require length of p be at least 20 days. In order to obtain daily data the follow sequence of values has been calculated: $(E(p_1), \dots, E(p_{128}))$ where p_1 represents the firsts 28 days (i.e. from 1 Feb. 2002 until 28 Feb. 2002), p_2 represents the seconds 28 days (i.e. from 28 Feb. 2002 until 1 Mar. 2002) and so on. Finally define Q_i be the mean quality for day i .

Now consider the index: $L_i = Q_i E(p_i)$. The goal is the study the behavior of $\ell_i = \text{logit}(L_i)$ conditioned to the process variables X_i . To avoid curse of dimensionality a drastic reduction of number of regressors was required.

Consequently an intensive exploratory data analysis has been performed in order to select few variables to implement the model. Different tools have been used such as graphical inspections with trellis graphics and other projection technique (see Insightful (2001)), step-wise selection techniques, and projection pursuit regression (see Friedmann and Stuetzle (1981), Huber (1985), Hall (1989)) as suggested in a previous study by Frederic and Catellani (2001).

Four variables, included the Z 's main flow (expressed in terms of kilos per hour), have been selected.

Figure 1: Scatter plot of the conditional mean $\mathbb{E}(\ell_{T+1}|x_{T+1}^{(j)})$ and the conditional standard error $\mathbb{V}^{1/2}(\ell_{T+1}|x_{T+1}^{(j)})$, with $j = 1, 10.000$.



4. Predictive Model

Let $\ell_{T+1} \in \mathbb{R}$ be the future value of variable ℓ for day $T + 1$, $x_{T+1} \in \mathbb{R}^4$ be the average of future values of the predictive variable for day $T + 1$, $\ell = (\ell_1, \dots, \ell_T) \in \mathbb{R}^T$ be the past T values of variable ℓ and $\mathbf{x} = (x_1, \dots, x_T) \in \mathbb{R}^{T \times 4}$ be the past values of variables xs .

The goal is to model:

$$f(\ell_{T+1}|x_{T+1}, \mathbf{x}, \ell) \propto \int_{\Theta} f(\ell_{T+1}|x_{T+1}, \theta) f(\ell|\mathbf{x}, \theta) dF(\theta) \quad (1)$$

where $f(\cdot|\cdot)$ represents the conditional probability density functions (the likelihood), $\theta \in \Theta$ a representation parameter and $F(\cdot)$ a suitable distribution on θ (the prior). Wide discussion on equation (1) can be find for example in De Groot (1970), Berger (1980), Gelman *et al.* (1995).

Since the previous exploratory data analysis gives evidence to some non-linear relation between x variables and ℓ and to presence of heteroscedasticity then a flexible heteroscedastic B-spline based additive model has been performed.

Let $N(\cdot, \cdot)$ be denoting the normal distribution, the likelihood model for $(\ell|x, \theta)$ is:

$$(\ell|x, \theta) \sim N(\theta_1(x), \theta_2(x)) \quad (2)$$

where

$$\theta_1(x) = B(x)\beta \quad (3)$$

$$\theta_2(x) = \exp\{B(x)\alpha\} \quad (4)$$

Here $B : \mathbb{R}^4 \rightarrow \mathbb{R}^{4M}$ is the set of B-spline, M is the number of B spline involved and $\beta \in \mathbb{R}^{4M}$ and $\alpha \in \mathbb{R}^{4M}$ the actual real representation parameters. A spline function is a continuous piece-wise polynomial function defined on a finite set of points called knots. A B-spline base is a finite set of some special easily computable spline (De Boor (1977), De

Boor (1978)). The space generated by such a base is wide enough to describe a large class of situations. Notice that if one denotes with symbol $\mathbb{E}(\cdot|\cdot)$ the conditional expected value then $\mathbb{E}(\ell|x, \theta_1) = \theta_1(x) = B(x)\beta$; that is equation (3) states that model (2) is an additive model (see for example Hastie and Tibshirani (1986), Buja *et al.* (1989)). Equation (4) states the standard deviation has been modelling as a function of x data.

5. Conclusions

Model described by equations (2), (3) and (4) has been implemented using the *WinBugs* package. Non-informative normal prior distributions have been used for parameters β and α . In order to explore the predictive function, defined in equation (1), 10.000 values $x_{T+1}^{(j)}, j = 1, \dots, 10.000$ have been randomly generated. Then, named $\mathbb{V}(\cdot|\cdot)^{1/2}$ the conditional standard deviation, the couples of points $(\mathbb{E}(\ell_{T+1}|x_{T+1}^{(j)}), \mathbb{V}(\ell_{T+1}|x_{T+1}^{(j)})^{1/2})$ have been plotted. Figure (1) shows such a plot. The shape of the cloud is clearly convex. An interesting result is that there exists a theoretical optimal value x_{T+1}^* , i.e. x_{T+1}^* is such that $\mathbb{E}(\ell_{T+1}|x_{T+1}^*) = \max$ and $\mathbb{V}(\ell_{T+1}|x_{T+1}^*)^{1/2} = \min$. Actually such a values are $\mathbb{E}(\ell_{T+1}|x_{T+1}^*) = 2,49$ and $\mathbb{V}(\ell_{T+1}|x_{T+1}^*)^{1/2} = 0,056$

References

- Bequette B.W. (1991) Non-linear control of chemical processes: A review, *Ind. Eng. Chem. Res.*, 30, 1391–1413.
- Berger J.O. (1980) *Statistical Decision Theory*, Springer-Verlag, New York.
- Buja A., Hastie T. and Tibshirani R. (1989) Linear smoothers and additive models, *Ann. of statist.*, 17, 453–455.
- Clarke D.W., Mohtadi C. and Tuffs P.S. (1987) Generalized predictive control-part i. the basic algorithm, *Automatica*, 23(2), 137–148.
- Coughanowr D. (1991) *Process Systems Analysis and Control*, McGraw-Hill, New York.
- De Boor C. (1977) Package for calculating with b-splines, *Siam J. numer. anal.*, 14, 441–472.
- De Boor C. (1978) *A practical guide to splines*, Spriger, Berlin.
- De Groot M.E. (1970) *Optimal Statistical Decisions*, McGraw-Hill, New York.
- Frederic P. and Catellani M. (2001) Statistical analysis of an industrial distillation process, *Statistica Applicata*, to appear.
- Friedmann J.B. and Stuetzle W. (1981) Projection pursuit regression, *JASA*, 76, 817–824.
- Garcia C.E., Prett D.M. and Morari M. (1989) Model predictive control: Theory and practice-a survey, *Automatica*, 25(3), 335–348.
- Gelman A., Carlin J.B., Stern H.S. and Rubin D.B. (1995) *Bayesian Data Analysis*, Chapman and Hall, London.
- Hall P. (1989) On projection pursuit regression, *Ann. Statist.*, 17, 573–588.
- Hastie T. and Tibshirani R. (1986) Generalized additive models, *Statistical Science*, 1, 297–318.
- Huber P.J. (1985) Projection pursuit, *Ann. Statist.*, 13, 435–475.
- Insightful (2001) *S-PLUS 6 for Windows User's Guide*, Insightful Corporation, Seattle, WA.
- Kirk R. (1998) *The Encyclopedia of Chemical Technology*, 4th ed. Wiley.