Computers and Chemical Engineering 80 (2015) 30–36

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Elastic net with Monte Carlo sampling for data-based modeling in biopharmaceutical manufacturing facilities

Kristen Severson[a], Jeremy G. VanAntwerp[a,b], Venkatesh Natarajan[c], Chris Antoniou[c], Jörg Thömmes[c], Richard D. Braatz[a,*]

[a] Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[b] Engineering Department, Calvin College, 3201 Burton SE, Grand Rapids, MI 49546, USA
[c] Biogen, 14 Cambridge Center, Cambridge, MA 02142, USA

ARTICLE INFO

Article history:
Received 5 March 2015
Received in revised form 26 April 2015
Accepted 5 May 2015
Available online 14 May 2015

Keywords:
Data-based modeling
Elastic net
Lasso
Monte Carlo methods
Biopharmaceutical manufacturing
Regularization methods

ABSTRACT

Biopharmaceutical manufacturing involves multiple process steps that can be challenging to model. Oftentimes, operating conditions are studied in bench-scale experiments and then fixed to specific values during full-scale operations. This procedure limits the opportunity to tune process variables to correct for the effects of disturbances. Generating process models has the potential to increase the flexibility and controllability of the biomanufacturing processes. This article proposes a statistical modeling methodology to predict the outputs of biopharmaceutical operations. This methodology addresses two important challenging characteristics typical of data collected in the biopharmaceutical industry: limited data availability and data heterogeneity. Motivated by the final aim of control, regularization methods, specifically the elastic net, are combined with sampling techniques similar to the bootstrap to develop mathematical models that use only a small number of input variables. This methodology is evaluated on an antibody manufacturing dataset.

1. Introduction

The U.S. biotechnology sector has had double-digit growth rates in recent years (IMARC Group, 2012). In 2012, sales of biologics were approximately $63.6 billion, with monoclonal antibodies (mAbs) representing the largest fraction of this market with approximately 39% of sales (Aggarwal, 2014). Modeling of the manufacturing process is one possible way to both support the growing biologics market as well as decrease costs *via* improved control and understanding of process operations. Modeling can play an important role in understanding, controlling, and optimizing the process steps used in these processes (Tziampazis and Sambanis, 1994). The U.S. Food and Drug Administration and International Conference on Harmonization recommend modeling in the development of biologics to estimate variability, provide process understanding, and establish a control strategy (U.S. Department of Health and Human Services, 2011; ICH, 2009).

Process modeling techniques can be grouped into two broad categories: first-principles and data-based. This article focuses on data-based modeling, which is more often applied in (bio)pharmaceutical manufacturing facilities. Data-based models have been applied to cell culture characterization (Mercier et al., 2013; Kirdar et al., 2007; Rathore et al., 2011), quality control (Roggo et al., 2007; Chen et al., 2011), process monitoring (Rathore et al., 2011; Read et al., 2010a,b; Bonné et al., 2013), and downstream operations (Rathore et al., 2011). A drawback of current data-based methods applied in the biopharmaceutical industry is that the models that are produced are not easily interpretable because they rely on subspaces that do not have direct physical meaning.

With this motivation, a successful biopharmaceutical model would achieve three goals: (1) model accuracy, (2) model simplicity, and (3) model interpretability. These aims have the caveat of using only a small amount of heterogeneous data, as data for biopharmaceutical manufacturing are typically both heterogeneous and relatively limited compared to most mature industries such as in chemicals, refining, petrochemicals, and pulp and paper.

One way to achieve these goals is through the identification of the input variables in the process that exhibit the largest effects on the output variables. It is common in the biopharmaceutical industry for a dataset to have more measurements, p, than observations, N. Most measurements are only taken once in a single batch moving through the production process and few replicates are performed due to time and cost constraints. The construction of predictive

[*] Corresponding author. Tel.: +1 617 253 3112; fax: +1 617 258 5042.
E-mail address: braatz@mit.edu (R.D. Braatz).

http://dx.doi.org/10.1016/j.compchemeng.2015.05.006
0098-1354/© 2015 Elsevier Ltd. All rights reserved.

models from such data sets can be made even more challenging because the collected data are typically highly correlated between batches, that is, the data sets are highly ill-conditioned. Regularization methods have been identified as possible approaches for such problems because of their ability to simultaneously handle input selection and model estimation (Pampuri et al., 2001).

This article first provides some background on regularization methods, specifically the lasso and elastic net. Modifications are then introduced to better handle small heterogeneous datasets. Finally, the methodology is evaluated for a manufacturing-scale process in the biopharmaceutical industry and the results are compared to other data-based modeling techniques used in the industry.

2. Background on regularization

The simplest form of regression finds a vector of weights, $\beta \in R^p$, that can be used to predict the scalar output y using the vector in inputs, $x \in R^p$. The basic approach to finding β is called *ordinary least squares* (OLS). The OLS problem is formulated to minimize the error:

$$\mathrm{Err}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2, \tag{1}$$

which has the solution

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty, \tag{2}$$

where $X \in R^{N\times p}$ when X^TX is invertible. Applying this method can lead to over-fitting of the model, especially as the number of input variables (p) grows large. Regularization techniques are one method to prevent over-fitting.

Lasso (Tibshirani, 1996), also known as ℓ_1 regularization, is an optimization formulation for parameter estimation that solves

$$\hat{\beta}_{\text{lasso}} = \underset{\beta,\,\beta_0}{arg\min}\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{3}$$

where N is the number of experiments, y_i is the ith scalar response, $x_i \in R^p$ is the data vector at observation i, λ is a nonnegative regularization parameter, β_0 is a scalar parameter, and $\beta \in R^p$ is a vector of model parameters. By adding the penalty term to the objective, the size of the coefficient vector is effectively constrained, which helps to prevent wild fluctuations of the coefficient vector that can be due to fitting noise in the data. The penalty is equivalent to the ℓ_1-norm on the coefficient vector, hence the name.

The lasso technique is useful to choose the subset of predictors (x_i) that exhibit the strongest effect on y because solutions to the lasso are sparse vectors, that is, the models only include a subset of the possible inputs ("dense" refers to models that include all possible inputs; these terms do not refer to the quality of data within each type of measurement). Because of the ℓ_1 constraint, solutions to the lasso can be thought of as lying on a vertex point of the feasible region, leading to certain coefficients to be exactly zero (Hastie et al., 2013; Rasmussen and Bro, 2012) (see Fig. 1 for a simple representation).

The elastic net (EN) (Zou and Hastie, 2005) is an optimization formulation for parameter estimation that is formulated as:

$$\hat{\beta}_{EN} = \underset{\beta_0,\beta}{arg\min}\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda P_\alpha(\beta), \tag{4}$$

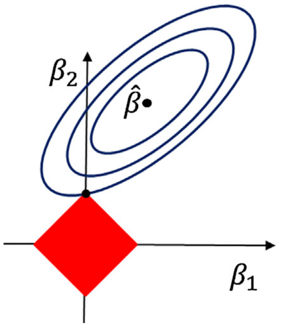

Fig. 1. Representation of parameter selection using the lasso considering the constrained optimization problem. The shaded region represents the constraint imposed by the penalty term and the ellipses are the contours of the least squares error function. The solution often lies on a vertex of the constraint, causing some parameters to be exactly zero. Figure based on (Hastie et al., 2013).

where

$$P_\alpha(\beta) = \sum_{j=1}^{p}\frac{1-\alpha}{2}\beta_j^2 + \alpha|\beta_j|, \tag{5}$$

N is the number of experiments, y_i is the ith scalar response, $x_i \in R^p$ is the data vector at observation i, λ is a nonnegative regularization parameter, β_0 is a scalar parameter, $\beta \in R^p$ is a vector of model parameters, and α is on the interval (0,1].

Although the elastic net is very similar to the lasso, there are some key differences. The EN is particularly useful when the number of predictors (p) is greater than the number of observations (N). If the lasso is applied to a data set where $p > N$, the solution is not unique. By adding a second term, the problem becomes convex even when $p > N$ (Zou and Hastie, 2005).

EN is also better at handling data where the inputs are correlated. Lasso is only able to select up to N predictors and will not reveal grouping relationships. Instead, the lasso will choose one of the correlated variables, and in highly correlated cases, can switch between variables in the set. However, the EN formulation uses a strictly convex penalty function and will guarantee that equal weighting is given to inputs that are identical (Zou and Hastie, 2005). Fig. 2 compares the penalty constraint contours for ridge regression which uses a quadratic penalty, lasso, and EN. EN can produce models that are sparse and can handle correlated data.

These points are illustrated here using a simple four-dimensional case study where x_1 and x_2 are specified then x_3 and x_4 are calculated to equal x_1 and x_2, respectively, with a small amount of random noise. The parameter traces in Fig. 3 show how EN groups the variables where the lasso uses one variable from each grouped set. The group is very robust to the selection of values of the penalty term λ. By including the grouped variables in the elastic net, the noise in the grouped measurements can be averaged in the calculation of the predictions produced by the model. Lasso selects a sparser model for a given value of λ, but at the cost of not being able

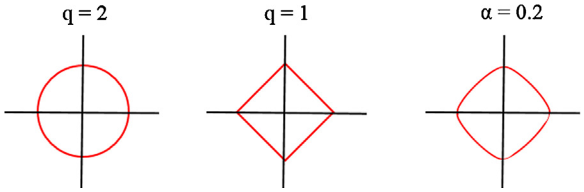

Fig. 2. Penalty constraint contours for constant values of $\sum_j |\beta_j|^q$ (left two images are for the ridge regression and lasso penalties) and $\sum_j \left(\alpha\beta_j^2 + (1-\alpha)|\beta_j|\right)$ (right-most image is for the elastic net penalty). Figure is based on (Hastie et al., 2013).