Computers and Chemical Engineering 73 (2015) 116–127

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

A combined first-principles and data-driven approach to model building

Alison Cozad[a], Nikolaos V. Sahinidis[a,b,*], David C. Miller[b]

[a] Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States
[b] U.S. Department of Energy, National Energy Technology Laboratory, Pittsburgh, PA 15236, United States

ARTICLE INFO

Article history:
Received 11 June 2014
Received in revised form 15 October 2014
Accepted 28 November 2014
Available online 8 December 2014

Keywords:
Regression
Surrogate models
Semi-infinite programming

ABSTRACT

We address a central theme of empirical model building: the incorporation of first-principles information in a data-driven model-building process. By enabling modelers to leverage all available information, regression models can be constructed using measured data along with theory-driven knowledge of response variable bounds, thermodynamic limitations, boundary conditions, and other aspects of system knowledge.

We expand the inclusion of regression constraints beyond intra-parameter relationships to relationships between combinations of predictors and response variables. Since the functional form of these constraints is more intuitive, they can be used to reveal hidden relationships between regression parameters that are not directly available to the modeler. First, we describe classes of *a priori* modeling constraints. Next, we propose a semi-infinite programming approach for the incorporation of these novel constraints. Finally, we detail several application areas and provide extensive computational results.

1. Introduction

Often, modelers must decide between (a) utilizing first-principles models, intuition, *etc.* or (b) constructing surrogate models using empirical data. We propose a combination of these two techniques that augments empirical modeling with first-principles information, intuition, and other *a priori* system characterization techniques to build accurate, physically realizable models. By doing this, we leverage the synergistic effects of empirical data, first-principles derivation, and intuition. Observed data points are often sampled at a premium, incurring costs associated with computational time, raw materials, and/or other resources. Frequently, additional insights provided by system knowledge, intuition, or the application of first-principles analysis are available without additional computational, financial, or other costly resource requirements. Knowledge of a less empirical nature, including limits on the response variables; known relationships between response and predictor variables; and relationships among responses, can be applied in conjunction with experimental data. For example, ensuring the nonnegativity of a modeled geometric length, enforcing a sum-to-one constraint on modeled chemical fractional compositions, and ensuring that derivative bounds obey thermodynamic principles are all practical applications of beneficial nonempirical insights.

We aim to build regression models (U):

$$\text{(U)} \quad \min_{\beta \in \mathcal{A}} \quad g(\beta; x_1, x_2, \ldots, x_N, z_1, z_2, \ldots, z_N)$$

that determine m regression parameters (coefficients) β that minimize a given loss function g (*e.g.*, squared error, regularized error, or an information criterion) over a set of original regression constraints $\mathcal{A}$ based on data points (x_i, z_i), $i = 1 \ldots N$. For conciseness, we will refer to $g(\beta; x_1, x_2, \ldots, x_N, z_1, z_2, \ldots, z_N)$ as $g(\beta)$.

To formally introduce insightful nonempirical information, we would like to enforce the following constraint on a regression problem:

$$\Omega(\mathcal{X}) := \left\{ \beta \in \mathbb{R}^m : f\left[x, \hat{z}(x; \beta)\right] \le 0, \quad x \in \mathcal{X} \right\} \tag{1}$$

where function f is a constraint in the space of the predictor(s) x and modeled response(s) $\hat{z}$, and $\mathcal{X}$ is a nonempty subset of $\mathbb{R}^n$. Eq. (1) can be used to reduce the feasible region $\mathcal{A}$ for any general regression analysis formulation: linear least squares, nonlinear least squares, regularized regression, best subset methods, and other characterization techniques. In fact, these constraints can be used alongside current gray-box or semi-physical modeling techniques (Nelles,

[*] Corresponding author at: Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States. Tel.: +1 412 2683338; fax: +1 412 268 7139.
E-mail address: sahinidis@cmu.edu (N.V. Sahinidis).

http://dx.doi.org/10.1016/j.compchemeng.2014.11.010
0098-1354/© 2014 Elsevier Ltd. All rights reserved.

2001; Pearson and Pottmann, 2000), where a balance between first principles knowledge and empirical data is desirable – typically, where model structure is chosen from system knowledge and parameters are selected to match sampled data.

By incorporating system knowledge beyond sampled data, we refine the feasible domain as the intersection of $\mathcal{A}$ and Ω and solve problem (C):

$$\text{(C)} \quad \min_{\beta \in \mathcal{A} \cap \Omega(\mathcal{X})} \quad g(\beta)$$

where Ω is defined over the domain $x \in \mathcal{X}$ while the original regression problem (U) exists in the space $\beta \in \mathcal{A}$.

Rao (1965) and Bard (1974) were the first to use *a priori* parameter relationships in regression through simple equality constraints. Recently, the use of such relationships has expanded to include inequality constraints in the space of the regression parameters, a case that arises more naturally in practice (Knopov and Korkhin, 2011). Inequality relationships between regression parameters have been applied to both linear and nonlinear least squares problems in the fields of statistics (Judge and Takayama, 1966; Liew, 1976), economics (Thompson, 1982; Rezk, 1976), and engineering (Gibbons and McDonald, 1999). Most notably, Korkhin has investigated the properties of simple parameter restrictions (Korkhin, 1985, 2002, 2005), nonlinear parameter restrictions (Korkhin, 1998, 1999), and, more recently, the formulation of inequality constraints with deterministic and stochastic right-hand sides (Korkhin, 2013).

Previous work employs *a priori* knowledge to reveal relationships between subsets of regression parameters that serve to restrict their range. To the best of our knowledge, there has been no investigation into the enforcement of *a priori* information that directly relates predictors to regressors. We aim to use these novel relationships between predictors and regressors to restrict the feasible region in the original problem space.

Since previous applications of constrained regression have been restricted to the parameter space β of the regression problem, these techniques are inherently specific to the functional form of the response. For example, consider a quadratic response model, $\hat{z}(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, and the *a priori* insight $\beta_1 \geq \beta_2$. If an exponential function, $\hat{z}(x) = \beta_0 + \beta_1 \exp(x)$, produces a more favorable fit, there is no standard way to translate constraints from the quadratic to the exponential model. On the other hand, enforcing a lower bound on the response, $\hat{z}(x) \geq 0 \quad \forall x$, rather than the β-space, produces a constraint that is independent of the model's functional form. Additionally, system insight in the x-domain may be more intuitive and readily available than knowledge of a unique and contrived regression model's functional form.

A complication arises from the realization that Eq. (1) is valid for the full problem space and, therefore, needs to be enforced for every point $x \in \mathcal{X}$, *i.e.*, at infinitely many points. Semi-infinite programming (SIP) problems are optimization models that have finitely many variables and infinitely many constraints (Reemtsen and Rückmann, 1998). These problem formulations are common in the fields of approximation theory, optimal control, and eigenvalue computations, among others. In each case, one or more parametric constraints result in one constraint for each value of an optimization parameter (in this case x) that varies within its given domain (Hettich and Kortanek, 1993; Reemtsen and Rückmann, 1998).

The first significant work on SIP, by John (1948), provides necessary and sufficient conditions for the solution to a semi-infinite program. Initially, SIP research focused on linear and convex nonlinear semi-infinite programming (Hettich and Kortanek, 1993; Reemtsen and Rückmann, 1998; Goberna and López, 2002). More recently, advances in global optimization, including BARON (Tawarmalani and Sahinidis, 2005), have made the solution of general nonconvex SIP problems more tractable (Chang and Sahinidis, 2011). In problem (C), the objective is often convex, as is the case for linear least squares regression. However, the feasible region – as we show in Section 4 – is generally nonlinear and nonconvex. The key to solving an SIP problem, independent of the solution method, is the optimization of $\max_{x \in \mathcal{X}} \; f\left[x, \hat{z}(x;\beta)\right]$ to locate the maximum violation. This subproblem is significant because $\beta \in \Omega(\mathcal{X})$ if and only if $\max_{x \in \mathcal{X}} f\left[x, \hat{z}(x;\beta)\right] \leq 0$ (Reemtsen and Görner, 1998).

To assess the benefits afforded by augmenting standard regression problems (U) with *a priori* information as in (C), we utilize the test platform ALAMO (Cozad et al., 2014). ALAMO is a software package designed to generate models that are as accurate and as simple as possible. This combination of accuracy and simplicity is well-suited for regression.

The remainder of the paper is organized as follows. In Section 2, we outline the modeling and sampling methods of the ALAMO test platform. We propose a methodology to solve problem (C) in its most general form in Section 3. In Section 4, we detail classes of applications of domain-constrained regression using our solution strategy: restricting individual and multiple responses, constraining response model derivatives, and expanding or contracting the enforcement domain. In Section 5, we demonstrate the efficacy of our approach using numerical examples. Next, in Section 6, we present extensive computational results demonstrating the effectiveness of the proposed methods. Finally, we offer conclusions in Section 7.

2. ALAMO

ALAMO is a learning software that identifies simple, accurate surrogate models using a minimal set of sample points from blackbox emulators such as experiments, simulations, and legacy code. ALAMO initially builds a low-complexity surrogate model using a best subset technique that leverages a mixed-integer programming formulation to consider a large number of potential functional forms. The model is subsequently tested, exploited, and improved through the use of derivative-free optimization solvers that adaptively sample new simulation or experimental points. For more information about ALAMO, see Cozad et al. (2014).

In this section, we detail relevant ALAMO model-building methods as applied to parametric regression. The functional form of a regression model is assumed to be unknown to ALAMO. Instead, ALAMO poses a simple set of basis functions, *e.g.*, x, x^2, $1/x$, $\log(x)$, and a constant term. Once a set of potential basis functions is collected, ALAMO attempts to construct the lowest complexity function that accurately models sample data. To do this, a mixed-integer quadratic program (MIQP) is solved to select basis functions for increasing model complexity. In a solution of the MIQP, the simple basis functions, $X_j(x), j \in \mathcal{B}$, are active when the corresponding binary variable $y_j = 1$ and inactive when $y_j = 0$. The size of the model, specified by a parameter T corresponding to the number of active binary variables, is increased until a goodness-of-fit measure, such as the corrected Akaike Information Criterion (Hurvich and Tsai, 1993), worsens with an increase in model size. As an example, using the list of basis functions given above, the MIQP is as follows:

$$\begin{array}{lll}
\text{(M)} \quad \min & g(\beta) = \displaystyle\sum_{i=1}^{N} \left(z_i - \left[\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \frac{1}{x} + \beta_4 \log(x) \right] \right)^2 & \\
\text{s.t.} & \beta_j^{\text{lo}} y_j \leq \beta_j \leq \beta_j^{\text{up}} y_j \qquad j = 0, \ldots, 4 & \\
& y_0 + y_1 + y_2 + y_3 + y_4 = T & \\
& y_j \in \{0, 1\} \qquad j = 0, \ldots, 4 &
\end{array}$$

While a typical basis set is often far larger, this simple example illustrates the form of the objective $g(\beta)$ and original constraint set $\beta \in \mathcal{A}$ before intersection with new *a priori* constraints $\Omega(\mathcal{X})$.