Analytica Chimica Acta 768 (2013) 57–68

Contents lists available at SciVerse ScienceDirect

Analytica Chimica Acta

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

Parsimonious and robust multivariate calibration with rational function Least Absolute Shrinkage and Selection Operator and rational function Elastic Net☆

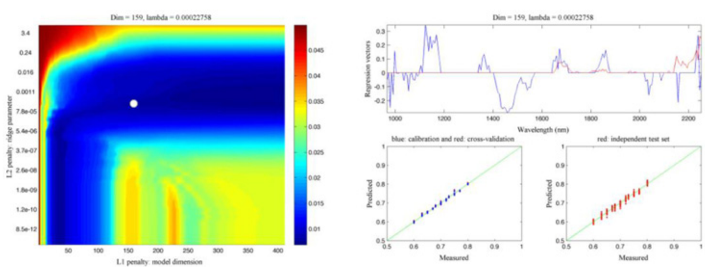
P. Teppola[a,*], V.-M. Taavitsainen[b]

[a] VTT Technical Research Centre of Finland, Photonic Devices and Measurement Solutions, P.O. Box 1199, FI-70211 Kuopio, Finland
[b] Helsinki Metropolia University of Applied Sciences, P.O. Box 4000, Bulevardi 31, FI-00079 Metropolia, Finland

HIGHLIGHTS

- A unique approach using rational functions and Elastic Net.
- Chemometric preprocessing not needed when using rational functions.
- Building parsimonious models in an automated way.
- Automated variable selection.
- A full continuum of feasible solutions with parsimony and/or smoothness.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 28 August 2012
Received in revised form 7 December 2012
Accepted 4 January 2013
Available online 7 February 2013

Keywords:
Chemometrics
Multivariate calibration
Rational function
LASSO
Elastic Net
RF-ENET
Multi-point
NIR

ABSTRACT

This paper presents new methods for multivariate calibration. A unique aspect is that this approach uses rational functions with either Least Absolute Shrinkage and Selection Operator (LASSO) or Elastic Net (ENET), and builds parsimonious models in an automated way via cross-validation. Rational function modeling provides robustness, as will be briefly demonstrated. Interestingly, rational function models are also flexible, in that occasionally they are reduced to ordinary linear models based on cross-validation. Thus, model complexity is not forced to take the form of rational functions.

Additional benefits arise from the use of LASSO and ENET. While LASSO uses only ℓ_1 norm on regression coefficients, ENET combines the best of both worlds by using ℓ_1 and ℓ_2 norms. The former (ℓ_1) provides variable selection while the latter (ℓ_2) handles collinearity via shrinkage of regression coefficients. Rational functions are highly collinear if full rank is used and, thus, not necessarily robust unless either ℓ_1 or ℓ_2 norm is used on the regression coefficients. The use of ℓ_1 norm allows for a more parsimonious model that can potentially be more robust. This is contrary to the use of a broadband spectrum that is likely to be contaminated at some point in the future by unknown spectral interferences. The real benefits seem to originate from the combination of rational functions and ENET. Note that LASSO solutions form a subset of ENET solutions and are thus included in ENET.

1. Introduction

The mainstream approach to multivariate calibration is to use a suitable means of one or two given spectral preprocessing techniques followed by an ordinary regression technique solving the multivariate calibration given by $\mathbf{y} = \beta_0 + \mathbf{X}\boldsymbol{\beta}$ where $\mathbf{y}$ is our response variable and $\mathbf{X}$ consists of measured spectral data. There is a real plethora of regression techniques, and most widely used are

☆ Paper presented at the XIII Conference on Chemometrics in Analytical Chemistry (CAC 2012), Budapest, Hungary, 25–29 June 2012.

* Corresponding author. Tel.: +358 20 722 2067; fax: +358 20 722 2080.
E-mail address: pekka.teppola@vtt.fi (P. Teppola).

0003-2670/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.aca.2013.01.005

Partial Least Squares (PLS) [1–4] in chemometrics and ridge regression [5] in statistics. These techniques use ℓ_2 norm on regression coefficients in order to take into account the collinearity in data. To be more accurate, PLS performs regression in a truncated basis [6] that indirectly leads to a shrinkage of regression coefficients in a ℓ_2 norm sense (compared to Ordinary Least Squares (OLS)).

Another common approach is to use techniques that carry out variable selection in order to robustify or simplify models. Among these techniques, we refer here to Least Absolute Shrinkage and Selection Operator (LASSO) [7]. This approach uses an ℓ_1 norm on regression coefficients in order to simplify models. Recently, a new approach called Elastic Net (ENET) [8] has been developed. This method is interesting in the sense that it combines both ℓ_1 and ℓ_2 norms. There is also an excellent paper from Kalivas with an overview on ℓ_1 and ℓ_2 norms and their applications to multivariate calibration and model maintenance [9].

A different approach has recently been developed by Taavitsainen, who has applied rational functions together with PLS and ridge regression [10]. This kind of rational function modeling adds flexibility in modeling by taking nonlinearities into account.

In this work, rational functions are combined with ENET, LASSO, ridge, and PLS, with a particular focus on the proposed rational function LASSO and ENET methods. In practical terms, the aim of this work is to illustrate and demonstrate the potential benefits of these new methods to real applications in multivariate calibration. Firstly, the robustness and flexibility of rational function modeling will be shown, and how rational functions can even replace the Standard Normal Variate (SNV) preprocessing [11] that is routinely used for correcting multiplicative spectral interferences and other optical path-length (OPL) changes. Secondly, we will demonstrate the added value of rational function LASSO (RF-LASSO) and ENET (RF-ENET) approaches and, in particular, how RF-ENET provides a full continuum of feasible solutions that make a balanced and flexible use of both parsimony (variable selection with ℓ_1 regularization) and smoothness (handling collinearity with ℓ_2 regularization).

2. Theory

2.1. Ridge regression (RR), Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ENET)

In 1970, Hoerl and Kennard introduced Ridge regression [5], which uses a regularization parameter λ_2 to control the inflation and general instability of least squares estimates. By doing this, regression coefficient estimates in ridge regression tend to be smoother and smaller in size, in addition to the fact that prediction estimates $\hat{\mathbf{y}}$ becoming more precise and stable. This is done by using an ℓ_2-penalty on regression coefficients.

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \tag{1}$$

where

$$\left\|\boldsymbol{\beta}\right\|_2^2 = \sum_{j=1}^{J} \beta_j^2$$

As well-known, the estimates of regression coefficients $\hat{\boldsymbol{\beta}}$ in ridge regression are computed as follows.

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = (\mathbf{X}^{\text{T}}\mathbf{X} + \lambda_2 \mathbf{I})^{-1}\mathbf{X}^{\text{T}}\mathbf{y} \tag{2}$$

The ridge regression usually provides smoothness to regression coefficients and takes care of any collinearity in data. It does not perform any variable selection, but rather uses all variables. On the other hand, Least Absolute Shrinkage and Selection Operator (LASSO) is used precisely for variable selection. LASSO uses an ℓ_1-penalty and continuously shrinks the smallest estimated regression coefficients towards zero. The number of zero-valued regression coefficients increases as a function of the regularization parameter λ_1.

$$\hat{\boldsymbol{\beta}}^{\text{Lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \tag{3}$$

where

$$\left\|\boldsymbol{\beta}\right\|_1 = \sum_{j=1}^{J} \left|\beta_j\right|$$

Various authors have listed a few issues in LASSO (Zou and Hastie [8], Zou and Zhang [12], Meinshausen and Bühlmann [13], and Zhao and Yu [14]). These shortcomings suggest that there may be some conditions in which LASSO becomes an inappropriate method for variable selection. For instance, in cases where there are more predictors than samples, LASSO is not well-defined unless the bound on the ℓ_1-norm of regression coefficients is smaller than a certain value as stated by Zou and Hastie. Other conditions are related to collinearity and pairwise correlations between predictors. It is for these reasons that Zou and Hastie have introduced Elastic Net (ENET) [8], combining both ℓ_1- and ℓ_2-penalties, and, by doing so, they try to fix the above shortcomings. The underlying (naïve) ENET model is given as follows.

$$\hat{\boldsymbol{\beta}}^{\text{Naïve elastic net}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \tag{4}$$

Based on empirical evidence, Zou and Hastie [8] call this a naïve ENET because it does not perform satisfactorily unless the solution is very close to either ridge regression or LASSO. This is due to the double shrinkage of applying first ℓ_2-penalty and then ℓ_1-penalty. In order to remedy this shortcoming of the naïve ENET, Zou and Hastie proposed a correction (*cf.* Eq. (7)) for the regression coefficient estimates. The solution that has been used in this work follows Eqs. (5)–(7)). Essentially, Eqs. (4) and(5) are the same. The solution uses augmented matrices $\mathbf{y}^*$ and $\mathbf{X}^*$.

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}^*}{\operatorname{argmin}} \|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \|\boldsymbol{\beta}^*\|_1 \tag{5}$$

where

$$\mathbf{y}^*_{(I+J)\times 1} = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix}$$

and

$$\mathbf{X}^*_{(I+J)\times J} = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_J \end{pmatrix}$$

This can be also expressed as an equivalent LASSO problem using augmented matrices as stated by Clemmensen et al. [15]

$$\frac{1}{\sqrt{1+\lambda_2}} \left(\mathbf{X}^T\mathbf{X} + \sqrt{\lambda_2}\mathbf{I}_J^T\mathbf{I}_J \right) \hat{\boldsymbol{\beta}}^* = \mathbf{X}^T\mathbf{y} \tag{6}$$

The correction of regression coefficient estimates is given by

$$\hat{\boldsymbol{\beta}}^{\text{Elastic net}} = \sqrt{1+\lambda_2}\,\hat{\boldsymbol{\beta}}^* \tag{7}$$

2.2. Ordinary Least Squares (OLS) and Partial Least Squares (PLS) regression

For Ordinary Least Squares (OLS), we refer to the paper by Hoerl and Kennard [5]. In addition, there is an excellent paper from Geladi and Kowalski [1] comparing (ordinary) Multiple Linear Regression (MLR, *i.e.* OLS), Principal Component Regression (PCR), and Partial