Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 189 (2017) 316–321

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Maximum likelihood estimation based regression for multivariate calibration

Lu Guo [a], Jiangtao Peng [a], Qiwei Xie [b,*]

[a] Faculty of Mathematics and Statistics, Hubei Key Laboratory of Applied Mathematics, Hubei University, Wuhan 430062, China
[b] Data Mining Lab., School of Economics and Management, Beijing University of Technology, Beijing 100124, China

ARTICLE INFO

Article history:
Received 13 January 2017
Received in revised form 3 August 2017
Accepted 9 August 2017
Available online 15 August 2017

Keywords:
Maximum likelihood estimation
Least-squares
Multivariate calibration
Regression

ABSTRACT

In this paper, we propose a maximum likelihood estimation based regression (MLER) model for multivariate calibration. The proposed MLER method seeks for the maximum likelihood estimation (MLE) solution of the least-squares problem, and it is much more robust to noise or outliers and accurate than the traditional least-squares method. An efficient iteratively reweighted least squares technique is proposed to solve the MLER model. As a result, our model can obtain accurate spectra–concentrate relations. Experimental results on three real near-infrared (NIR) spectra data sets demonstrate that the proposed MLER model is much more efficacious and effective than state-of-the-art partial least squares (PLS) methods.

1. Introduction

As an effective tool for chemometrics applications, multivariate calibration (MVC) aims to reveal the intrinsic quantitative relations between the spectra and corresponding concentrations by means of a regression model. According to Beer-Lambert law, the regression relationship between absorbance and concentration usually follows a linear regression model. The multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) models [1,2] are usually used to describe the linear spectra–concentration relation.

However, the intrinsic regression process in the above-mentioned regression techniques is actually least-squares (LS) regression [3], which is sensitive to noise and outliers [4]. The quadratic loss function in the LS model measures data and noise on the same norm scale and does not differentiate noise from data [5]. In the presence of noise or outliers, the estimated regression coefficient may also reflect the effect of noisy samples (outliers). For spectroscopic data, the noises and outliers are usually encountered during the data measurement and acquisition, such as instrument noise, environmental noise, non-representative sampling and so on [6,7]. These noises or outliers will make the least-squares-based regression coefficient vector distorted [8].

In order to eliminate the effect of noise or outliers, many robust regression methods have been proposed for the MVC, such as robust SIMPLS (RSIMPLS) [9] and partial robust regression (PRM) [10]. RSIMPLS is a robust version of the classic SIMPLS method by using two robust steps: a robust covariance (or robust scores) estimation and a robust regression. PRM combines M-regression and PLS regression by employing two kinds of weights to handle the vertical outliers and leverage points in a latent variables regression model. There are also many other algorithms that try to improve the robustness of PLS by modifying the original PLS weight, such as power PLS (PPLS) [11] and sparse matrix transform based PLS (SMTPLS) [12]. In PPLS, the PLS weight is computed by taking powers of the $\mathbf{y}-\mathbf{X}$ correlations and $\mathbf{X}$ standard deviations, which neutralizes the influence of dominance of irrelevant $\mathbf{X}$-variance and spurious $\mathbf{y}$-correlations [11]. In SMTPLS, a sparse matrix transform technique is first used to decorrelate the observation data, then the PLS loading weight is computed in the decorrelated data space by least squares regression. The SMT decorrelation operation can alleviate the effect of correlated variables to the least squares computation of PLS weight.

Rather than making robust estimation for PLS in the latent variable space, we preform the robust regression in the original data space and propose a robust maximum likelihood estimation based regression (MLER) model in this paper. Inspired by the robust regression theory [13,14], in the original LS model, we replace the objective term of quadratic loss function to an MLE-like estimator, which minimizes a function of the fitting residuals. The minimization problem can be easily transformed into an iteratively reweighted LS problem,

* Corresponding author.
E-mail address: qwxiecasia@163.com (Q. Xie).

http://dx.doi.org/10.1016/j.saa.2017.08.020
1386-1425/

where a reasonable weight function is designed for the spectroscopic regression. The proposed MLER model utilizes the MLE principle to robust the LS estimator, which can obtain an accurate spectra–concentrate relation.

2. Maximum Likelihood Estimation Based Regression

Let $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_n]^\mathrm{T}$ be the $n \times p$ data matrix with each sample $\mathbf{x}_i \in \mathcal{R}^p$, and $\mathbf{y} = [y_1, \dots, y_n]^\mathrm{T}$ be the $n \times 1$ response vector. In general, the relationship between absorbance $\mathbf{X}$ and concentration $\mathbf{y}$ follows Beer-Lambert law, which yields to a linear regression relation. The multiple linear regression (MLR) model is usually used to describe the spectra–concentration relation as follows:

$$\min_{\beta} \| \mathbf{y} - \mathbf{X}\beta \|^2, \tag{1}$$

where β is the regression coefficient computed based on the least-squares criterion. The least-squares method requires the fitted residual $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$ has symmetric and continuous probability distribution. However, in practice, this may be not true, especially when outliers (abnormal data or noise) occur in spectroscopic data. In the presence of outliers, the least-squares model Eq. (1) will largely affected because the quadratic loss function in the LS model does not differentiate noise from data [5]. In order to eliminate the effects of noise or outliers and obtain a robust model, a maximum likelihood estimation-based regression (MLER) model is proposed.

Inspired by the robust regression theory [13,14], the least-squares objective function in Eq. (1) can be designed as an MLE-like estimator, which is associated with the distribution of the coding residuals.

Denote the residual $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta = [e_1, e_2, \dots, e_n]^\mathrm{T}$ with each element $e_i = y_i - \mathbf{x}_i^\mathrm{T}\beta, \quad i = 1, 2, \dots, n$. Assume that $e_1, e_2, \dots, e_n$ are independently and identically distributed according to some probability density function (PDF) $f_\theta(e_i)$, where θ denotes the parameter set that characterizes the distribution. Hence, we can get the likelihood of the estimator as follow:

$$L_\theta(e_1, e_2, \dots, e_n) = \prod_{i=1}^{n} f_\theta(e_i). \tag{2}$$

The MLE aims to maximize this likelihood function or, equivalently, minimize the objective function

$$-\ln L_\theta = \sum_{i=1}^{n} \rho_\theta(e_i), \tag{3}$$

where $\rho_\theta(e_i) = -\ln f_\theta(e_i)$.

The MLE of regression coefficient vector β can be formulated as the following minimization problem:

$$J(\beta) = \min_{\beta} \sum_{i=1}^{n} \rho_\theta \left(y_i - \mathbf{x}_i^\mathrm{T}\beta \right). \tag{4}$$

In Eq. (4), the PDF f_θ is unknown. We should provide some prior knowledge on the distribution f_θ for solving the optimization problem Eq. (4). In general, we can assume that the unknown PDF $f_\theta(e_i)$ is symmetric, and $f_\theta(e_i) < f_\theta(e_j)$ if $|e_i| > |e_j|$. So $\rho_\theta(e_i)$ has the following properties: $\rho_\theta(0)$ is the global minimal of $\rho_\theta(e_i)$; $\rho_\theta(e_i) = \rho_\theta(-e_i)$; $\rho_\theta(e_i) > \rho_\theta(e_j)$ if $|e_i| > |e_j|$. Without loss of generality, we let $\rho_\theta(0) = 0$. With these general assumptions of ρ_θ, in the following, we will transform the minimization problem in Eq. (4) into an iteratively reweighted least squares problem, which can derive a weight with clear physical meaning, i.e., noisy points will have low weight values.

Let $F_{\boldsymbol\theta}(\mathbf{e}) = \sum_{i=1}^{n} \rho_{\boldsymbol\theta}(e_i)$. In order to produce a convex optimization problem, we approximate $F_{\boldsymbol\theta}(\mathbf{e})$ by its first order Taylor expansion in the neighborhood of $\mathbf{e}_0$ as:

$$\tilde{F}_{\boldsymbol\theta}(\mathbf{e}) = F_{\boldsymbol\theta}(\mathbf{e}_0) + (\mathbf{e} - \mathbf{e}_0)^\mathrm{T} F'_{\boldsymbol\theta}(\mathbf{e}_0) + \frac{1}{2}(\mathbf{e} - \mathbf{e}_0)^\mathrm{T} U (\mathbf{e} - \mathbf{e}_0), \tag{5}$$

where $F'_{\boldsymbol\theta}(\mathbf{e})$ is the derivative of $F_{\boldsymbol\theta}(\mathbf{e})$. Denote by $\rho'_{\boldsymbol\theta}$ the derivative of $\rho_{\boldsymbol\theta}$, then $F'_{\boldsymbol\theta}(\mathbf{e}_0) = [\rho'_{\boldsymbol\theta}(e_{0,1}); \rho'_{\boldsymbol\theta}(e_{0,2}); \cdots; \rho'_{\boldsymbol\theta}(e_{0,n})]$, where $e_{0,i}$ is the i-th element of $\mathbf{e}_0$. The third term on the right hand of Eq. (5) is the high order residual term, and U is a diagonal matrix for that the elements in $\mathbf{e}$ are independent and there is no cross term between e_i and e_j, $i \neq j$, in $F_{\boldsymbol\theta}(\mathbf{e})$.

Since $F_{\boldsymbol\theta}(\mathbf{e})$ reaches its minimal value (i.e., 0) at $\mathbf{e} = \mathbf{0}$, we also require that $\tilde{F}_{\boldsymbol\theta}(\mathbf{e})$ has its minimal value at $\mathbf{e} = 0$. Taking the derivative of $\tilde{F}_{\boldsymbol\theta}(\mathbf{e})$ with respect to $\mathbf{e}$, it gets:

$$\tilde{F}'_{\boldsymbol\theta}(\mathbf{e}) = \tilde{F}'_{\boldsymbol\theta}(\mathbf{e}_0) + U(\mathbf{e} - \mathbf{e}_0).$$

Let $\tilde{F}'_{\boldsymbol\theta}(\mathbf{0}) = \mathbf{0}$, we have the diagonal element of U as:

$$U_{i,i} = u_{\boldsymbol\theta}(e_{0,i}) = \frac{\rho'_{\boldsymbol\theta}(e_{0,i})}{e_{0,i}}. \tag{6}$$

According to the properties of $\rho_{\boldsymbol\theta}(e_i)$, $\rho'_{\boldsymbol\theta}(e_i)$ will have the same sign as e_i. So each $U_{i,i}$ is a non-negative scalar. Then $\tilde{F}_{\boldsymbol\theta}(\mathbf{e})$ can be rewritten as $\tilde{F}_{\boldsymbol\theta}(\mathbf{e}) = \frac{1}{2} \| U^{1/2}\mathbf{e} \|_2^2 + b$, where b is a scalar value determined by $\mathbf{e}_0$. Since $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$, the MLER model in Eq. (4) can be approximated by

$$J(\beta, U) = \min_{\beta, U} \| U^{1/2}(\mathbf{y} - \mathbf{X}\beta) \|_2^2 . \tag{7}$$

The MLE weight matrix U needs to be estimated using Eq. (6), so Eq. (7) is a local approximation of the MLER in Eq. (4) at $\mathbf{e}_0$, and the minimization procedure of MLER can be transformed into an iteratively reweighted least square problem with U being updated using the residuals in previous iteration via Eq. (6). Each $U_{i,i}$ is a non-negative scalar, so the weighted least squares in each iteration is a convex problem, which could be solved easily. Therefore, we can obtain the solution of Eq. (7) as below:

$$\beta = (\mathbf{X}^\mathrm{T} U \mathbf{X})^{-1} (\mathbf{X}^\mathrm{T} U \mathbf{y}). \tag{8}$$

In the practice, when the number of variables is larger than the number of samples, the least-squares solution in Eq. (8) is usually unstable. To improve the stability of the solution, a regularized least-squares solution can be used:

$$\beta = (\mathbf{X}^\mathrm{T} U \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\mathrm{T} U \mathbf{y}, \tag{9}$$

where λ is a regularization parameter and $\mathbf{I}$ is an identity matrix.

Since U is a diagonal matrix, its element $U_{i,i}$ (i.e., $u_{\boldsymbol\theta}(e_i)$) is the MLE weight assigned to each e_i. Intuitively, the noises should have low weight values. Thus, with Eq. (7), the determination of distribution $\rho_{\boldsymbol\theta}$ is now transformed into the determination of MLE weight U. Considering that the logistic function has properties similar to the