Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Error Covariance Penalized Regression: A novel multivariate model combining penalized regression with multivariate error structure

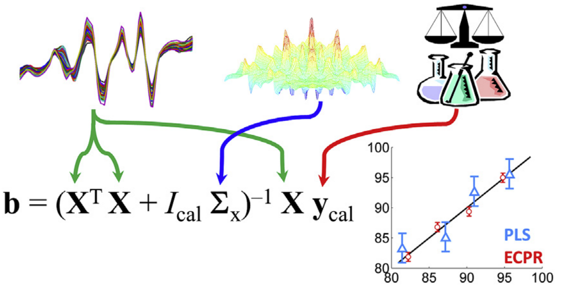
Franco Allegrini [a], Jez W.B. Braga [a, b], Alessandro C.O. Moreira [b, c], Alejandro C. Olivieri [a, *]

[a] Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Suipacha 531, Rosario, S2002LRK, Argentina
[b] Laboratório de Automação, Quimiometria e Química Ambiental, Instituto de Química, Universidade de Brasília, CEP 70904-970, Brasília, DF, Brazil
[c] Laboratório de Produtos Florestais, Serviço Florestal Brasileiro, CEP 70818-900, Brasília, DF, Brazil

HIGHLIGHTS

- A new multivariate penalized regression model was developed.
- It includes information on the error covariance matrix.
- The performance is better than classical multivariate models.
- It is extremely simple from the computational viewpoint.

GRAPHICAL ABSTRACT


$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + I_{\mathrm{cal}}\,\Sigma_{\mathrm{x}})^{-1}\,\mathbf{X}\,\mathbf{y}_{\mathrm{cal}}$
100
95
90
85
80
80
85
90
95
100
PLS
ECPR

ARTICLE INFO

Article history:
Received 31 October 2017
Received in revised form
23 January 2018
Accepted 1 February 2018
Available online xxx

Keywords:
Multivariate calibration
Penalized regression
Error covariance matrix

ABSTRACT

A new multivariate regression model, named Error Covariance Penalized Regression (ECPR) is presented. Following a penalized regression strategy, the proposed model incorporates information about the measurement error structure of the system, using the error covariance matrix (ECM) as a penalization term. Results are reported from both simulations and experimental data based on replicate mid and near infrared (MIR and NIR) spectral measurements. The results for ECPR are better under non-iid conditions when compared with traditional first-order multivariate methods such as ridge regression (RR), principal component regression (PCR) and partial least-squares regression (PLS).

1. Introduction

When developing multivariate calibration models, the structure of the instrumental noise is usually assumed to be independently and identically distributed (iid), although this appears to be the exception rather than the rule [1]. The noise information is contained in the error covariance matrix (ECM), a square matrix with diagonal elements that are the variances at each instrumental sensor, and the off-diagonal elements measure the covariance or degree of correlation of the noise at different sensors [1–3]. When the noise is not iid, incorporating this information into the multivariate calibration models improves the prediction ability in comparison with classical counterparts [4–8]. Various procedures have been proposed for the experimental estimation of the ECM from replicate sample measurements [1,2].

Inverse multivariate calibration is today the model of choice for

[*] Corresponding author.
E-mail address: olivieri@iquir-conicet.gov.ar (A.C. Olivieri).

https://doi.org/10.1016/j.aca.2018.02.002
0003-2670/© 2018 Elsevier B.V. All rights reserved.

a myriad of applications, which are mainly concentrated in the field of near infrared (NIR) spectroscopy [9], although additional multivariate signals including optical spectra [UV–visible, mid infrared (MIR)] or other sources (nuclear magnetic resonance, chromatography, electrochemical traces, etc.) are available for similar purposes. In these inverse models, the calibration phase solves the inverse problem $\mathbf{y}_{cal} = \mathbf{X}\ \mathbf{b}$, where $\mathbf{y}_{cal}$ is the vector of calibration concentrations for the analyte of interest, $\mathbf{X}$ is the matrix of full calibration spectra and $\mathbf{b}$ the vector of regression coefficients. The bottle-neck step when solving this model is the inversion of the matrix product ($\mathbf{X}^T\mathbf{X}$), (the superscript 'T' implies transposition). Because $\mathbf{X}$ usually contains considerably more wavelengths than samples, the product ($\mathbf{X}^T\mathbf{X}$) is singular, and different strategies are employed to tackle this issue [10].

Maximum likelihood principal component regression (MLPCR) is an inverse calibration model which considers the noise structure in its formulation [4,6,11]. It can be described as a PCR model where: (1) maximum likelihood principal component analysis (MLPCA) [12,13], rather than PCA, is employed in the decomposition of the calibration data matrix, and (2) a maximum likelihood projection, rather than an orthogonal projection into the PCA subspace, is used in the prediction step [4]. As other classical inverse models, MLPCR solves the inverse calibration problem by reducing the dimensionality of the full-spectral data by a projection onto a handful of latent variables. MLPCR has been shown to outperform classical PCR and partial least-squares (PLS) when the noise deviates from the iid paradigm [4,5].

An apparently unrelated family of inverse multivariate calibration models includes the so-called penalized regression. They solve the inverse calibration problem by adding a penalized term to the non-invertible matrix product ($\mathbf{X}^T\mathbf{X}$). The term can be as simple as a small multiple of a unit matrix, as in ridge regression (RR) [14,15], or may include matrices of different complexity as in Tikhonov's regularization [16–18]. These latter models have found important applications in calibration transfer and maintenance [19,20]. Penalized regression has also been successfully applied in other relevant fields, such as: (1) fluorescence microscopy, to achieve the blind deconvolution and detection of the origin of emission sources and to increase image resolution [21,22], (2) time resolved spectroscopy [23] and (3) determination of diffusion coefficients in pulsed gradient spin echo NMR data [24].

In the present report, we describe a simple inverse multivariate model incorporating the noise structure information, based on penalized regression using the error covariance matrix for penalization. It is extremely simple in computational terms (only a single programming line is needed for estimating the vector of regression coefficients, see Appendix), it requires no latent variables estimation, and provides analytical prediction results which are comparable to those furnished by MLPCR. When non-iid noise is present, it outperforms the classical PCR, PLS and RR models, as confirmed in a variety of simulated data sets with controlled noise properties. To the best of our knowledge, this is the first time experimental data sets including replicates from both calibration and independent validation samples are discussed. The latter sets were analyzed using the proposed model, with similar results to those obtained for the simulations. The improved analytical ability can be explained by the model inclusion of error covariance matrices estimated from extensive replicate analysis, which showed that the experimental noise was not iid.

2. Theory

2.1. Model

ECPR is based on the specific inverse least-squares model developed by Brown [25], although the latter can only be applied when the pure spectra of all sample constituents are known (unlike most inverse multivariate models). In the present case, Brown's model is adapted to the usual situation where the spectra are measured for mixtures, and only the concentration of a single analyte is known. Analyte prediction (y) proceeds through the usual expression:

$$y = \mathbf{x}\ \mathbf{b} \tag{1}$$

where $\mathbf{x}$ is the test sample spectrum, and $\mathbf{b}$ is given by:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X} + I\ \Sigma_x)^{-1}\ \mathbf{X}^T\ \mathbf{y}_{cal} \tag{2}$$

where I is the number of calibration samples and $\mathbf{\Sigma}_x$ is the error covariance matrix characterizing the structure of the instrumental noise. The latter can be estimated as described in Refs. [1,2]. Equation (2) is the solution of the minimization of an objective function which represents a trade-off between (squared) calibration error and prediction uncertainty, i.e., $\mathbf{b}$ is estimated as:

$$\mathbf{b} = \arg\min(||\mathbf{X}\ \mathbf{b} - \mathbf{y}_{cal}\ ||^2\ /\ I + \mathbf{b}^T\ \Sigma_x\ \mathbf{b}) \tag{3}$$

where || || indicates the Euclidean norm.

In the present report, $\mathbf{\Sigma}_x$ is proposed to be the (simulated or experimental) error covariance matrix for the test sample. In the simulations, the matrix inversion in equation (2) is not problematic; however, in the experimental cases $\mathbf{\Sigma}_x$ may be singular or near-singular. This issue can be solved by incorporating an additional penalized term in equation (2):

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X} + I\ \Sigma_x + \lambda\mathbf{I})^{-1}\ \mathbf{X}^T\ \mathbf{y}_{cal} \tag{4}$$

where $\mathbf{I}$ is an appropriately dimensioned unit matrix and λ a tunable parameter, which can be easily estimated by cross-validation [26] or by resorting to the so-called L-curve [27,28], as in RR-type methods. Recently, a somewhat related penalized regression model has been discussed in the context of calibration maintenance, using sample residuals as penalization term instead of the complete error covariance matrix [29]. In the present study, the root mean square error in cross-validation (RMSECV) values vs. the number of latent variables were considered for optimizing the PLS, PCR and MLPCR models, and RMSECV vs. λ for optimizing the RR and ECPR models.

Even when ECPR is extremely simple from the computational point of view (it requires a single program line, see Appendix), it efficiently takes into account the noise structure, with results which are comparable with seemingly more complex models such as MLPCR. An advantage of ECPR might be the fact that it does not employ latent variables: it is well-known that the estimation of the number of components for PCR and PLS by cross-validation is sometimes dependent of the specific procedure employed and on rather subjective judgments that can lead to different results. However, ECPR requires the estimation the λ parameter, and, as all ML-based models, the collection of replicate spectra to estimate $\mathbf{\Sigma}_x$. The performance of ECPR will be compared in this report with classical PCR/PLS, MLPCR and RR. The theory of these latter models can be found in the relevant literature.

2.2. Simulated data

Simulated spectra for three compounds (one analyte of interest and two interferents) were generated using Gaussian curves, varying the following parameters: (1) degree of selectivity, (2) noise type and (3) noise magnitude. Calibration and test