Research Article

JOURNAL OF CHEMOMETRICS

Received: 21 July 2009, Revised: 20 November 2009, Accepted: 21 November 2009, Published online in Wiley Interscience: 2010

(www.interscience.wiley.com) DOI: 10.1002/cem.1277

Multivariate calibration leverages and spectral *F*-ratios via the filter factor representation

Erik Andries[a,b]* and John H. Kalivas[c]

Diagnostics are fundamental to multivariate calibration (MC). Two common diagnostics are leverages and spectral *F*-ratios and these have been formulated for many MC methods such as partial least square (PLS), principal component regression (PCR) and classical least squares (CLS). While these are some of the most common methods of calibration in analytical chemistry, ridge regression is also common place and yet spectral *F*-ratios have not been developed for it. Noting that ridge regression is a form of Tikhonov regularization (TR) and using the unifying filter factor representation for MC, this paper develops the filter factor form of leverages and spectral *F*-ratios. The approach is applied to a spectral data set to demonstrate computational speed-up advantages and ease of implementation for the filter factor representation. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: multivariate calibration; leverage; spectral *F*-ratio; PCR, PLS; Tikhonov regularization

1. INTRODUCTION

Multivariate calibration (MC) methods in chemometrics are powerful tools used by industry for quality and process control. In spectroscopy, for example, MC methods construct a mathematical model that relates concentration or some other sample property to the spectra of known calibration samples. The constructed model is then applied to the spectrum of an unknown sample to predict its concentration or other property. Common MC methods include classical least squares (CLS), principal component regression (PCR), Tikhonov regularization (TR, also known as ridge regression under certain conditions) and partial least squares (PLS) [1–4]. Diagnostics are important to all MC methods. For example, if outliers are identified as being present and are removed from the calibration set, greater calibration accuracy can be achieved. However, if outliers are not removed, then they can have a strong influence on the estimation of the model parameters. During prediction, diagnostics are used to detect outlier samples whose spectra are sufficiently different from the calibration spectra. This paper focuses on two common outlier diagnostic measures used for spectral data—leverages and spectral *F*-ratios [5,6]. It should be noted that these two diagnostic measures have pitfalls, e.g. lack of statistical robustness and the curse of high dimensionality when samples sparsely populate a high-dimensional wavelength space [7–14]. Nonetheless, leverages and spectral *F*-ratios are still widely used for outlier detection and are in fact recommended by the American Society for Testing Materials [5]. Hence, the goal of this paper is to present a unified computational framework for this important class of widely used diagnostic measures.

In the chemometrics literature, there is a preference for MC methods such as PCR and PLS that are projection-based whereby high-dimensional calibration spectra are projected onto a lower-dimensional subspace. Ideally, the resulting subspace simultaneously includes and excludes components unharmed by and dominated by noise, respectively, and calibration, prediction and diagnostics are performed in this lower-dimensional subspace. However, calibration, prediction and diagnostics depend not only on the number of dimensions kept but also on the type of basis vectors used to form the subspace. For example, PCR, as its name implies, operates in the subspace spanned by the principal component directions formed by the singular value decomposition (SVD) of the calibration data. PLS, on the other hand, operates in the space spanned by the *Krylov subspace* [15–21]. When either subspace spans the original calibration space, then the calculation of diagnostics will yield the same result. However, when the subspace is strictly a lower-dimensional subspace (a subspace of lower numerical rank), then there will be differences in the diagnostics since the model spaces used to represent the reconstructed data come from different bases.

Ultimately, one has to choose which set of basis vectors to work with. In this paper, we do not advocate one basis set over another. Our interest is in the following: if we choose the basis vectors derived from the SVD, then regression and the diagnostics, leverage and spectral *F*-ratios, can be subsumed under the umbrella of the *filter factor representation*. Moreover, one can easily extend these two diagnostics to other MC methods that are not projection-based such as TR. While a non-filter factor

* *Correspondence to: Erik Andries, The University of New Mexico, The Center for Advanced Research Computing, MSC01 1190, 1 University of New Mexico, Albuquerque, NM 87131-0001, E-mail: andriese@hpc.unm.edu*

a *Erik Andries*
Center for Advanced Research Computing, University of New Mexico, Albuquerque, NM 87106, USA

b *Erik Andries*
Department of Mathematics, Central New Mexico Community College, Albuquerque, NM 87106, USA

c *John H. Kalivas*
Department of Chemistry, Campus Box 8023, Idaho State University, Pocatello, ID 83209, USA

representation of leverage for TR has been developed [22–24], to date, spectral *F*-ratio formulas for TR have not been developed in the literature. The goal of this paper is to unify the computation of leverages and spectral *F*-ratios for all MC methods under a single framework. It is well known that MC methods such as CLS, PCR, TR and PLS are functionally equivalent in how they perform regression using the filter factor representation [17,19,25–28]. However, our focus is to show that these MC methods are also functionally equivalent in how they compute leverages and spectral *F*-ratios as well.

The paper is organized as follows: In Section 2, we discuss the SVD and the need for regularized least squares other than CLS. In Section 3, we discuss how regression-based MC methods are unified under the filter factor representation and Section 4 details how leverages and spectral *F*-ratios can be similarly unified via this representation. Section 5 gives the numerical results and Section 6 states the conclusion.

All vectors are column vectors, unless otherwise indicated. Italicized lowercase symbols represent scalars (x) while upright and bold symbols represent vectors ($\mathbf{x}$). Matrices are indicated with upright boldface uppercase characters ($\mathbf{X}$). The superscripted symbols $^{\mathrm{T}}$ and † indicate the transpose and pseudoinverse, respectively, of a vector or matrix. A vector of n ones or zeros is indicated by $\mathbf{1}_n$ and $\mathbf{0}_n$, respectively, while $\mathbf{I}_n$ represents the identity matrix of dimension n. An m by n matrix of zeros will be indicated by $\mathbf{0}_{m,n}$. An m by p matrix $\mathbf{A}$ can be formed by concatenating p column vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ of dimension m such that $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p]$. The ıth row of the matrix $\mathbf{A}$ will be denoted by $\mathbf{a}_{,i}$. Diagonal matrices will be denoted using the MATLAB-like notation 'diag', e.g. $\mathbf{I}_n = \mathrm{diag}(\mathbf{1}_n) = \mathrm{diag}([1, 1, \ldots, 1]^{\mathrm{T}})$. The m by n matrix $\mathbf{X}$ containing the spectroscopic data consists of m samples stacked row-wise such that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^{\mathrm{T}}$ where the jth sample, denoted as the column vector $\mathbf{x}_j = [x_{j1}, x_{j2}, \ldots, x_{jn}]^{\mathrm{T}}$, is an element of the Euclidean n-space $\mathbb{R}^n$. The vector $\mathbf{y} = [y_1, y_2, \ldots, y_m]^{\mathrm{T}}$ represents the response variables such that y_j is the response variable associated with the jth sample $\mathbf{x}_j$. In the context of analyte concentration prediction from near-infrared spectra, $\mathbf{X}$ represents calibration data of n absorbance measurements across m samples and $\mathbf{y}$ represents m non-negative analyte concentration measurements.

2. SVD AND REGULARIZATION

It is standard procedure to mean center (column-wise) the calibration data:

$$\mathbf{X} := \mathbf{X} - \mathbf{1}_m\bar{\mathbf{x}}^{\mathrm{T}}, \quad \mathbf{y} := \mathbf{y} - \mathbf{1}_m\bar{y} \tag{1}$$

where $\bar{\mathbf{x}} = \frac{1}{m}(\mathbf{X}^{\mathrm{T}}\mathbf{1}_m)$ and $\bar{y} = \frac{1}{m}(\mathbf{y}^{\mathrm{T}}\mathbf{1}_m)$ denote the mean spectrum and mean response, respectively. Prediction for a future mean-centered spectrum ($\mathbf{z} := \mathbf{z} - \bar{\mathbf{x}}$) can then be written as $\hat{y} = \bar{y} + \mathbf{z}^{\mathrm{T}}\mathbf{b}$ where $\mathbf{b}$ is the regression vector. Unless otherwise indicated, it is assumed that $\mathbf{X}$ and $\mathbf{y}$ have already been mean-centered in a column-wise fashion via Equation(1).

2.1. Singular value decomposition

When solving the linear system $\mathbf{Xb} = \mathbf{y}$, CLS decomposition techniques such as LU or QR factorization will be numerically suspect if $\mathbf{X}$ is ill-conditioned. The unreliability of computing the solution vector $\mathbf{b}$ is best understood if we express $\mathbf{b}$ in terms of the *full* SVD of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}_{\mathrm{full}}\mathbf{S}_{\mathrm{full}}\mathbf{V}_{\mathrm{full}}^{\mathrm{T}}$$

where $\mathbf{U}_{\mathrm{full}}$ is an m by m orthonormal matrix, $\mathbf{V}_{\mathrm{full}}$ is an n by n orthonormal matrix and $\mathbf{S}_{\mathrm{full}}$ is an m by n diagonal matrix of *singular values*. The non-zero singular values are arranged in decreasing order such that $s_1 \geq s_2 \geq \cdots s_r \geq 0$ where $r \leq \min(m-1, n)$ is the rank of $\mathbf{X}$. The matrices $\mathbf{U}_{\mathrm{full}}$, $\mathbf{S}_{\mathrm{full}}$ and $\mathbf{V}_{\mathrm{full}}$ associated with the full SVD of $\mathbf{X}$ can also be represented as

$$\mathbf{U}_{\mathrm{full}} = [\mathbf{U}, \mathbf{U}_{\mathrm{null}}], \quad \mathbf{S}_{\mathrm{full}} = \begin{bmatrix} \mathbf{S} & \mathbf{0}_{\mathrm{r,m-r}} \\ \mathbf{0}_{\mathrm{m-r,r}} & \mathbf{0}_{\mathrm{m-r,m-r}} \end{bmatrix}, \quad \mathbf{V}_{\mathrm{full}} = [\mathbf{V}, \mathbf{V}_{\mathrm{null}}]^{\mathrm{T}}$$

where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$ denotes the first r columns of $\mathbf{U}_{\mathrm{full}}$, $\mathbf{U}_{\mathrm{null}} = [\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m]$ denotes the last $m-r$ columns of $\mathbf{U}_{\mathrm{full}}$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ denotes the first r columns of $\mathbf{V}_{\mathrm{full}}$ and $\mathbf{V}_{\mathrm{null}} = [\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n]$ denotes the last $n-r$ columns of $\mathbf{V}_{\mathrm{full}}$. Note that when $m > n$ and $r = n$, then $\mathbf{S}_{\mathrm{full}} = [\mathbf{S}, \mathbf{0}_{\mathrm{r,m-r}}]^{\mathrm{T}}$ and $\mathbf{V}_{\mathrm{full}} = \mathbf{V}$. The columns of the matrix $\mathbf{U}$ form an orthonormal basis for the range of $\mathbf{X}$ (the space spanned by the columns of $\mathbf{X}$) while the columns of the matrix $\mathbf{V}$ form an orthonormal basis for the range of $\mathbf{X}^{\mathrm{T}}$ (the space by the rows or calibration samples of $\mathbf{X}$ except that each row is the transpose of a vector in $\mathbb{R}^n$). The columns of the matrices $\mathbf{V}_{\mathrm{null}}$ and $\mathbf{U}_{\mathrm{null}}$ form an orthonormal basis for the nullspace of $\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}$, respectively, and will play a non-trivial role in the computation of leverages and spectral *F*-ratios. Since $\mathbf{U}_{\mathrm{null}}$ and $\mathbf{V}_{\mathrm{null}}$ are multiplied by the zeros in $\mathbf{S}_{\mathrm{full}}$, the SVD of $\mathbf{X}$ is typically written in its *reduced* form

$$\mathbf{X} = \mathbf{U}_{\mathrm{full}}\mathbf{S}_{\mathrm{full}}\mathbf{V}_{\mathrm{full}}^{\mathrm{T}} = \mathbf{USV}^{\mathrm{T}}$$

The reduced SVD is often used to compute the generalized inverse or the Moore–Penrose pseudoinverse of $\mathbf{X}$ where $\mathbf{X}^{\dagger} = \mathbf{VS}^{-1}\mathbf{U}^{\mathrm{T}}$.

Principal component analysis (PCA) is the SVD applied to mean-centered data. PCA is among the most widely used techniques in statistics, chemometrics, data analysis and data mining. PCA also forms the basis of many regression and machine learning methods. Computationally, PCA amounts to the low-rank approximation of a matrix containing the spectra being analyzed.

2.2. Classical least squares and SVD

Using the Moore–Penrose pseudoinverse of $\mathbf{X}$, the CLS solution can be written as

$$\mathbf{b}_{\mathrm{CLS}} = \mathbf{b}_{\mathrm{CLS}} = \mathbf{X}^{\dagger}\mathbf{y} = \mathbf{VS}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{y} = \mathbf{V}\boldsymbol{\alpha} = \sum_{i=1}^{r}\alpha_i\mathbf{v}_i \tag{2}$$

$$\boldsymbol{\alpha} = \mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{y} = [\alpha_1, \ldots, \alpha_r]^{\mathrm{T}}, \quad \alpha_i = \frac{\mathbf{u}_i^{\mathrm{T}}\mathbf{y}}{s_i} \tag{3}$$

Due to the orthonormality of $\mathbf{V}$, we have the following relation: $||\mathbf{b}_{\mathrm{CLS}}||_2^2 = ||\mathbf{V}\boldsymbol{\alpha}||_2^2 = ||\boldsymbol{\alpha}||_2^2$. Phenomenologically, the smallest singular values are typically associated with spurious noise inherent in the data and singular vectors $\mathbf{v}_i$ that are rough and highly oscillatory in profile [17]. The division of $\mathbf{u}_i^{\mathrm{T}}\mathbf{y}$ by a small singular value s_i unduly amplifies the size of the vector-norm of $\mathbf{b}_{\mathrm{CLS}}$ and, as a consequence, the corresponding basis vector $\mathbf{v}_i$ dominates the CLS solution. Statistically, large-norm CLS solutions have low *bias* and high *variance*. From a prediction point of view, large-norm

solutions are associated with overfitting. Statistical techniques such as PCR, PLS and TR shrink the size of the regression coefficients in **b** in order to obtain a small-norm solution.

3. REGULARIZATION STRATEGIES

The term *regularization* generally refers to incorporating *a priori* information into the regression model in order to stabilize the regression vector against the effects of noise and to sift out a spectroscopically plausible solution. Regularization, in this paper, refers to obtaining small-norm solutions and can generally be divided into two broad classes: projection and penalty methods.

In projection methods, the solution vector is restricted to span a lower-dimensional subspace $\mathcal{S}$ consisting of k dimensions:

$$\min_{\mathbf{b}\in\mathcal{S}} \phi(\mathbf{b}), \quad \phi(\mathbf{b}) = ||\mathbf{Xb} - \mathbf{y}||_2^2 \tag{4}$$

The regularization of projection methods depends upon the number of subspace dimensions kept. If $k = r$ (i.e. full rank), then the original CLS solution is recovered. However, it is hoped that $k << r$ with $\mathcal{S}$ containing only 'pure' components uncontaminated by noise. The two most common chemometric techniques, PCR and PLS, are projection methods. In contrast to projection methods, penalty methods filter out unwanted noise components by adding a penalty term to the least squares problem:

$$\min \phi(\mathbf{b}), \quad \phi(\mathbf{b}) = ||\mathbf{Xb} - \mathbf{y}||_2^2 + \lambda^d ||\mathbf{Mb} - \mathbf{g}||_d^d \tag{5}$$

The type and amount of regularization is controlled by changing the parameters d, λ, **M** and **g**. For example, the regression problem known as 'the LASSO' corresponds to a L_1-norm penalty formulation of TR with the following parameters: $d = 1$, $\mathbf{M} = \mathbf{I}_n$ and $\mathbf{g} = \mathbf{0}$ [29]. However, we will concern ourselves with the TR form that is characterized by a L_2-norm penalty term ($d = 2$). TR with $\mathbf{M} = \mathbf{I}_n$ and $\mathbf{g} = \mathbf{0}$ is often referred to as the *standard* TR (STR) problem whereas the $\mathbf{M} \neq \mathbf{I}_n$ case is referred to as the *generalized* TR (GTR) problem [3]. Note that in the statistics literature, STR is more commonly known as *ridge regression* [4].

Both STR or GTR can be thought of as the CLS solution applied to the following augmented minimization problem:

$$\min \phi(\mathbf{b}), \quad \phi(\mathbf{b}) = ||\mathbf{X}_\lambda \mathbf{b} - \mathbf{y}_\lambda||_2^2 \quad \text{where}$$
$$\mathbf{X}_\lambda = \begin{bmatrix} \mathbf{X} \\ \lambda\mathbf{M} \end{bmatrix}, \quad \mathbf{y}_\lambda = \begin{bmatrix} \mathbf{y} \\ \lambda\mathbf{g} \end{bmatrix}$$

The above equation essentially approximates the following equality-constrained least squares problem: solve $\mathbf{Xb} = \mathbf{y}$ subject to the equality constraints $\mathbf{Mb} = \mathbf{g}$. In the seminal text by Lawson and Hanson [30], three methods are specified for solving equality-constrained least squares problems. The first two methods involve hard modeling approaches which rigorously satisfy both $\mathbf{Xb} = \mathbf{y}$ and $\mathbf{Mb} = \mathbf{g}$ to a high degree of accuracy. The third method is a soft modeling approach and is precisely the penalty approach of GTR. Instead of trying to rigorously satisfy both $\mathbf{Xb} = \mathbf{y}$ and $\mathbf{Mb} = \mathbf{g}$, GTR seeks a trade off between minimizing the residual norm $||\mathbf{Xb} - \mathbf{y}||_2$ and the equality constraint norm $||\mathbf{Mb} - \mathbf{g}||_2$. Note that when $\mathbf{M} = \mathbf{I}_n$ in Equation (5), each element of the vector $\mathbf{Mb} - \mathbf{g}$ is equally weighted. Alternatively, λ in Equation (5) can be replaced with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ if one seeks to give a different weight to each element of $\mathbf{Mb} - \mathbf{g}$. This approach is often referred to as *generalized ridge regression*. In the chemometrics literature, a thorough statistical analysis of ridge and generalized ridge regression can be found in References [31,32].

We now briefly discuss how these projection- and penalty-based regularization strategies can be accommodated by the filter factor representation.

3.1. Regularization via filter factors

PCR, PLS and TR are often treated as disparate regression algorithms. However, these methods can also be unified under a broader class of methods called *filter factor* methods. These methods aim to shrink the size of solution norm $||\mathbf{b}||_2$ in Equation (2) by pre-multiplying $\mathbf{S}^{-1}$ by a diagonal filter factor matrix $\mathbf{F} = \text{diag}(\mathbf{f})$ where $\mathbf{f} = [f_1, f_2, \ldots, f_r]^\mathrm{T}$ so as to create a *regularized* inverse $\mathbf{X}^\# = \mathbf{VFS}^{-1}\mathbf{U}^\mathrm{T}$. The regularized solution $\mathbf{b}_{\text{REG}}$ can then be expressed as a re-weighted version of the CLS solution:

$$\mathbf{b}_{\text{REG}} = \mathbf{X}^\#\mathbf{y} = \mathbf{VFS}^{-1}\mathbf{U}^\mathrm{T}\mathbf{y} = \mathbf{VF}\boldsymbol{\alpha} = \sum_{i=1}^{r} f_i \alpha_i \mathbf{v}_i \tag{6}$$

The matrix **F** also re-weights the square of the CLS solution norm $||\mathbf{b}_{\text{CLS}}||_2^2 = ||\boldsymbol{\alpha}||_2^2$ such that

$$||\mathbf{b}_{\text{REG}}||_2^2 = ||\mathbf{VF}\boldsymbol{\alpha}||_2^2 = ||\mathbf{F}\boldsymbol{\alpha}||_2^2 = \sum_{i=1}^{r} f_i^2 \left(\frac{\mathbf{u}_i^\mathrm{T}\mathbf{y}}{s_i}\right)^2 \tag{7}$$

If f_i is zero or small in magnitude, then the ıth dimension of the SVD basis plays no or little role in constructing the solution vector. If the filter factors f_i are small for large i, then the regularization scheme employed resembles a low-pass filter in signal processing where the solution is smoothed by removing rough singular vectors.

The filter factors associated with the common regression techniques of PCR, STR and PLS are well known [17,19,25–28] and are summarized below:

$$f_i = \begin{cases} 1 \text{ if } i \leq k \leq r \text{ and } 0 \text{ otherwise} & \text{PCR} \\ \dfrac{s_i^2}{s_i^2 + \lambda^2} & \text{STR} \\ 1 - \mathcal{R}_k(s_i^2) & \text{PLS} \end{cases} \tag{8}$$

where $i = 1, \ldots, r$. For PCR, the regularization parameter is k—the dimension of the subspace spanned by the first k singular vectors. As a consequence, $f_i = 1$ for the first k terms of Equation (6) and $f_i = 0$ for the remaining terms. For STR, the regularization parameter is $\lambda > 0$ and all r terms of Equation (6) are kept but the filter factors are weighted such that $f_1 > f_2 > \cdots > f_r$ where $f_i \in [0, 1]$. Effectively, terms in the SVD expansion associated with large i (small singular values) are damped to a greater degree than terms associated with small i (large singular values). For PLS, the regularization parameter is k—the dimension of the *Krylov subspace* denoted by

$$\mathcal{K}_k(\mathbf{G}, \mathbf{d}) = \text{span}\{\mathbf{d}, \mathbf{Gd}, \ldots, \mathbf{G}^{k-1}\mathbf{d}\} \tag{9}$$

where $\mathbf{G} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{d} = \mathbf{X}^T\mathbf{y}$. The term $\mathcal{R}_k(\theta)$ in Equation (8) is the value of the *Ritz polynomial* of degree k evaluated at $\theta = s_i^2$. On average, the filter factors f_i decay from 1 to 0 as i increases

but the oscillatory nature of the Ritz polynomial can sometimes cause f_i to take values outside of the interval [0, 1]. A complete derivation of the filter factors for each of these methods is found in the Appendix.

3.2. Native basis set for PLS

It is important to note the association between the common regression techniques of PCR, STR and PLS and their *native* set of basis vectors used to span the calibration space of the spectra. For PCR and STR, the basis vectors formed by SVD are the standard choice. However, the basis vectors associated with PLS are the Krylov subspace vectors in Equation (9) [18]. These vectors (i.e. $\mathbf{G}^i\mathbf{d} = (\mathbf{X}^\mathrm{T}\mathbf{X})^i\mathbf{X}^\mathrm{T}\mathbf{y}$, $i = 0, \ldots, k-1$) rapidly become linearly dependent in finite precision arithmetic as i increases since successive vectors point more and more in the direction of the dominant singular vector $\mathbf{v}_1$ [34]. After $\mathbf{v}_1$, the next set of preferential directions are, on average, $\mathbf{v}_2$, $\mathbf{v}_3$ and so on. Since the basis vectors of the Krylov subspace point preferentially in the directions of the dominant singular vectors and since these vectors capture most of the space spanned by the calibration spectra, PLS tends to converge faster than PCR in terms of the number of basis vectors used to form the regression vector. This is a computational advantage over PCR and is one of the reasons why it is used to solve large-scale linear systems. Yet, regression vector computation does not imply subspace superiority in terms of prediction. Using simulation studies of complex chemical mixtures containing large number of components, Reference [35] was unable to show any evidence to support that PLS performs 'better' than PCR for prediction. It is also important to note that the Krylov subspace vectors are not actually used to compute the regression vector since they do not form a good basis set. For numerical stability, the Lanczos bidiagonalization procedure (LBP) is often used to orthonormalize the Krylov subspace [17,21,33].

The LBP, at step k, creates a lower bidiagonal matrix $\tilde{\mathbf{R}}_k$ of dimension $(k+1) \times k$ and two orthogonal matrices $\tilde{\mathbf{U}}_{k+1}$ and $\tilde{\mathbf{V}}_k$ such that

$$\mathbf{X} \approx \tilde{\mathbf{U}}_{k+1}\tilde{\mathbf{R}}_k\tilde{\mathbf{V}}_k^\mathrm{T} \tag{10}$$

where the matrix $\tilde{\mathbf{V}}_k$ forms an orthonormal basis for the vectors in Equation (9). When applied to the normal equations, the Lanczos bidiagonalization of $\mathbf{X}$ and $\mathbf{y}$ becomes the Lanczos *tridiagonalization* of $\mathbf{X}^\mathrm{T}\mathbf{X}$ and $\mathbf{X}^\mathrm{T}\mathbf{y}$. Instead of creating a bidiagonal matrix, the LBP creates a bidiagonal matrix that is also symmetric, i.e. a tridiagonal matrix $\tilde{\mathbf{T}}_k$ such that $\tilde{\mathbf{T}}_k = \tilde{\mathbf{R}}_k^\mathrm{T}\tilde{\mathbf{R}}_k$ and $\mathbf{X}^\mathrm{T}\mathbf{X} = \tilde{\mathbf{V}}_k\tilde{\mathbf{T}}_k\tilde{\mathbf{V}}_k$. The zeros of the Ritz polynomial $\mathcal{R}_k(\theta)$ in Equation (8)—called the Ritz values—also have the remarkable property of being the eigenvalues of $\tilde{\mathbf{T}}_k$ [15,17,21]. This connection between the Ritz values and the eigenvalues of $\tilde{\mathbf{T}}_k$ yields the following crucial insight: the regularizing action of PLS is determined by how well the Ritz values approximate s_i^2 (the singular values of $\mathbf{X}^\mathrm{T}\mathbf{X}$). If a Ritz value has converged to s_i^2, then $\mathcal{R}_k(s_i^2) = 0$ and $f_i^k = 1$. However, if s_i^2 lies between Ritz values that have not converged, then $f_i^k = 1 - \mathcal{R}_k(s_i^2)$ can oscillate in value since $\mathcal{R}_k(\theta)$ is a polynomial of degree k. Details concerning the oscillatory nature of the PLS-based filter factors can be found in References [17,19]. Since the convergence behavior of the Ritz values is not known *a priori*, the filter factors for PLS are likewise not known *a priori*.

When solving $\mathbf{Xb} = \mathbf{y}$ using the first k iterations of the LBP, the pseudoinverse of the right-hand side of Equation (10) is used to get

$$\mathbf{b} = \tilde{\mathbf{V}}_k\tilde{\mathbf{R}}_k^{-1}\tilde{\mathbf{U}}_{k+1}^\mathrm{T}\mathbf{y}$$

In this Krylov subspace setting, PLS is a projection method. However, when the basis vectors are changed from $\tilde{\mathbf{V}}_k$ to the singular vectors in $\mathbf{V}$, then PLS behaves not like a projection method but more like STR in the sense that the filter factors $f_i = 1 - \mathcal{R}_k(s_i^2)$ are not binary and decay, on average, from 1 to 0.

PLS is only as good as the basis vectors used to represent the regression vector. As the number of LBP iterations increases, numerical round-off error can degrade the orthonormal integrity of the matrices $\tilde{\mathbf{U}}_{k+1}$ and $\tilde{\mathbf{V}}_k$ in Equation (10). In the numerical analysis literature, it is common practice to *re-orthonormalize* $\tilde{\mathbf{U}}_{k+1}$ and $\tilde{\mathbf{V}}_k$ using modified Gram–Schmidt or Householder re-orthonormalization procedures [17,36]. However, these re-orthonormalization procedures are largely absent from chemometric PLS software packages and have only recently been brought to the attention of the chemometric community [37]. If one uses *many* PLS factors to model data, then the computation of the regression vector $\mathbf{b}$ and diagnostic merits such as leverages and spectral F-ratios can be compromised when Krylov subspace approaches are used without re-orthogonalization since the matrices $\tilde{\mathbf{U}}_k$ and $\tilde{\mathbf{V}}_k$ will be no longer orthogonal due to catastrophic round-off error [17,36–38].

3.3. Non-standard filter factor procedures

Standard filtering procedures invoke a smoothness assumption: the low frequency terms in the SVD expansion associated with the largest singular values are the most important and should be kept, while the high frequency terms corresponding to the smallest singular values should be filtered out or damped. Although keeping the dominant SVD expansion terms has been the primary tool for data representation, it need not be the most effective strategy for prediction. For example, in principle component selection, k dimensions of the SVD subspace are kept but they are not necessarily the first k dimensions. These k dimensions are chosen on the basis on how well they are correlated with or how well they predict the response variable $\mathbf{y}$. If $\mathcal{F}$ denotes the set of the chosen k dimensions, then the corresponding filter factors are $f_i = 1$ if $i \in \mathcal{F}$ and $f_i = 0$ otherwise. In principle, the filter factors could be arbitrary. In practice, however, the filter factors are such that the sum $\sum_{i=1}^{r} f_i$ generally ranges between 0 and the rank of the data r. This sum is often referred to as the *effective numerical rank* or the *effective degrees of freedom* [17,26]. Regardless of which filter factors are used, the overarching goal is to obtain the relation $||\mathbf{b}_{\mathrm{REG}}||_2 < ||\mathbf{b}_{\mathrm{CLS}}||_2$ that yields good prediction for an unseen sample.

4. LEVERAGE AND SPECTRAL *F*-RATIOS VIA A FILTER FACTOR REPRESENTATION

Many statistical diagnostic measures are available with MC methods. We concentrate on perhaps the two most commonly used ones—leverage and the spectral F-ratio [5,6]. One use of these diagnostic merits is for outlier detection [5,6]. In the discussion that follows, the reader is reminded that both the calibration spectra $\mathbf{X}$ and the validation spectrum $\mathbf{z}$ have already been mean-centered relative to $\bar{\mathbf{x}}$ as described in Equation (1).

4.1. Leverage

In a high-dimensional setting, the validation spectrum $\mathbf{z}$ will not likely lie in the space spanned by the calibration spectra. Hence, we have to consider the portions of $\mathbf{z}$ that lie and do not lie in the space spanned by the calibration set. As a result, it will be convenient to express the validation sample $\mathbf{z}$ as the sum of two vectors lying in different orthogonal subspaces $\mathcal{S}_1$ (the range of $\mathbf{X}^T$ or the space spanned by the calibration spectra) and $\mathcal{S}_2$ (the nullspace of $\mathbf{X}$ which is orthogonal $\mathcal{S}_1$):

$$\mathbf{z} = \mathbf{V}_{\text{full}}\mathbf{w} = [\mathbf{V}, \mathbf{V}_{\text{null}}]\begin{bmatrix}\mathbf{w}\\ \mathbf{w}_{\text{null}}\end{bmatrix} = \mathbf{V}\mathbf{w} + \mathbf{V}_{\text{null}}\mathbf{w}_{\text{null}} = \mathbf{z}_1 + \mathbf{z}_2 \quad (11)$$

where $\mathbf{w}$ and $\mathbf{w}_{\text{null}}$ are r- and $(n-r)$-dimensional, respectively, with $\mathbf{z}_1 = \mathbf{V}\mathbf{w} \in \mathcal{S}_1$ and $\mathbf{z}_2 = \mathbf{V}_{\text{null}}\mathbf{w}_{\text{null}} \in \mathcal{S}_2$. The vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ of $\mathbf{z}$ also correspond to the following orthogonal projections:

$$\mathbf{z}_1 = \mathbf{P}\mathbf{z} = (\mathbf{V}\mathbf{V}^{\mathsf{T}})\mathbf{z}, \quad \mathbf{z}_2 = (\mathbf{I}_n - \mathbf{P})\mathbf{z} = (\mathbf{V}_{\text{null}}\mathbf{V}_{\text{null}}^{\mathsf{T}})\mathbf{z} \quad (12)$$

where $\mathbf{P} = \mathbf{V}\mathbf{V}^{\mathsf{T}}$ orthogonally projects any vector in $\mathbb{R}^n$ onto $\mathcal{S}_1$. Since $\mathbf{z}_1$ and $\mathbf{z}_2$ belong to different orthogonal subspaces, we have the Pythagorean relation: $||\mathbf{z}||_2^2 = ||\mathbf{z}_1||_2^2 + ||\mathbf{z}_2||_2^2$. The decomposition of $\mathbf{z}$ into its orthogonal components will be shown to significantly impact how the spectral F-ratio merit can detect the 'outlyingness' of $\mathbf{z}$.

4.1.1. Mahalanobis distance

If the distribution of the calibration samples is spherical, then the Euclidean distance could reliably be used to determine whether the validation sample $\mathbf{z}$ is an outlier relative to the calibration samples. However, if the distribution is not spherical but ellipsoidal, then we would expect that the probability of $\mathbf{z}$ being an outlier to depend not only on the distance from the origin but also on the direction of the major and minor axes. That is, all points lying on the surface of an ellipsoid will have the same distance. This distance between $\mathbf{z}$ and the origin is known as the Mahalanobis distance, and will be indicated by $d(\mathbf{z})$. In the case of mean-centered data, $d(\mathbf{z})$ is classically expressed as

$$d(\mathbf{z}) = \sqrt{\mathbf{z}^{\mathsf{T}}\mathbf{C}^{\dagger}\mathbf{z}}, \quad \mathbf{C} = \frac{\mathbf{X}^{\mathsf{T}}\mathbf{X}}{m-1}$$

where $\mathbf{C}$ is the sample covariance matrix. A contour plot of increasing values for $d(\mathbf{z})$ would result in series of 'concentric' ellipsoids increasing in size about the origin.

4.1.2. Leverage via classical least squares

The diagnostic merit known as leverage is related to the Mahalanobis distance. Since $\mathbf{X}$ has already been mean-centered (as opposed to the case where the intercept variable is included such that $\mathbf{X}_{\text{pad}} = [X, \mathbf{1}_n]$), leverage is simply the square of the Mahalanobis distance divided by $m-1$

$$h(\mathbf{z}) = \frac{d^2(\mathbf{z})}{m-1} = \mathbf{z}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{\dagger}\mathbf{z} \quad (13)$$

where $h(\mathbf{z})$ denotes the leverage associated with the sample $\mathbf{z}$. Note that

$$\mathbf{z}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{\dagger}\mathbf{z} = ||(\mathbf{X}^{\dagger})^{\mathsf{T}}\mathbf{z}||_2 = ||\mathbf{U}\mathbf{S}^{-1}\mathbf{V}^{\mathsf{T}}\mathbf{z}||_2 = ||\mathbf{S}^{-1}\mathbf{V}^{\mathsf{T}}\mathbf{z}||_2 = ||\boldsymbol{\beta}||_2$$

where $\boldsymbol{\beta} = \mathbf{S}^{-1}\mathbf{V}^{\mathsf{T}}\mathbf{z} = [\beta_1, \beta_2, \ldots, \beta_r]^{\mathsf{T}}$ and $\beta_i = \mathbf{v}_i^{\mathsf{T}}\mathbf{z}/s_i$. As a result, Equation (13) can be rewritten as the square of the two-norm of $\boldsymbol{\beta}$:

$$h(\mathbf{z}) = \mathbf{z}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{\dagger}\mathbf{z} = ||\boldsymbol{\beta}||_2^2 = \sum_{i=1}^{r}\left(\frac{\mathbf{v}_i^{\mathsf{T}}\mathbf{z}}{s_i}\right)^2 \quad (14)$$

Equation (14) also identifies the meaningful leverage constituents: $h(\mathbf{z})$ is being inflated by squared components of $\boldsymbol{\beta}$ associated with the division of small singular values, just as the solution norm $||\mathbf{b}||_2$ was similarly inflated in Equation (3).

The expression in Equation (14) is also related to the 'hat' matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}$ that orthogonally projects $\mathbf{y}$ onto a subspace of $\mathbb{R}^m$ (the range of $\mathbf{X}$). When $m \geq n$ and $\mathbf{X}$ has full rank ($r = n$), then the predicted values $\hat{\mathbf{y}}$ for the calibration set become

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{\text{CLS}} = \mathbf{X}\mathbf{X}^{\dagger}\mathbf{y} = \mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} = \mathbf{U}\mathbf{U}^{\mathsf{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$$

The leverage for the jth calibration sample is then the value of the jth diagonal element of $\mathbf{H}$ since $h(\mathbf{x}_j) = \mathbf{x}_j^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_j = \mathbf{H}_{jj}$. If we interpret $\mathbf{U}$ and $\mathbf{S}\mathbf{V}^{\mathsf{T}}$ as the score and loading matrices of $\mathbf{X}$, respectively, then the leverage value for the jth sample simply becomes the sum of its corresponding squared score values:

$$h(\mathbf{x}_j) = \mathbf{H}_{jj} = (\mathbf{U}\mathbf{U}^{\mathsf{T}})_{jj} = ||\mathbf{u}_{:,j}^{\mathsf{T}}||_2^2 = \sum_{i=1}^{r} u_{ji}^2$$

Note that $\boldsymbol{\beta} = \mathbf{S}^{-1}\mathbf{V}^{\mathsf{T}}\mathbf{x}_j = \mathbf{u}_{:,j}^{\mathsf{T}}$ when $\mathbf{x}_j = \mathbf{V}\mathbf{S}\mathbf{u}_{:,j}^{\mathsf{T}}$. Hence, when $\mathbf{z} = \mathbf{x}_j$, we can regard the components of the vector $\boldsymbol{\beta}$ as the score values for the jth sample.

4.1.3. Leverage via regularized least squares

When the pseudoinverse of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ in Equation (13) is replaced with the regularized inverse of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, Equation (14) becomes

$$h(\mathbf{z}) = \mathbf{z}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{\#}\mathbf{z} = ||(\mathbf{X}^{\#})^{\mathsf{T}}\mathbf{z}||_2 = ||\mathbf{F}\boldsymbol{\beta}||_2^2 = \sum_{i=1}^{r} f_i^2\left(\frac{\mathbf{v}_i^{\mathsf{T}}\mathbf{z}}{s_i}\right)^2 \quad (15)$$

Again, the filter factors are damping out terms associated with small singular values. Geometrically, the filter factors preferentially shrink the ellipsoid in the direction of the basis vectors of $\mathbf{V}$ associated with small singular values, causing the ellipsoid to become less oblate. When the validation sample is replaced with the jth calibration sample, Equation (15) simplifies to

$$h(\mathbf{x}_j) = \mathbf{x}_j^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{\#}\mathbf{x}_j = ||\mathbf{F}\boldsymbol{\beta}||_2^2 = ||\mathbf{F}\mathbf{u}_{:,j}^{\mathsf{T}}||_2 = \sum_{i=1}^{r} f_i^2 u_{ji}^2$$

The hat matrix $\mathbf{H}$ similarly undergoes a 'filtered' modification

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{\text{REG}} = \mathbf{X}\mathbf{X}^{\#} = \mathbf{U}\mathbf{F}\mathbf{U}^{\mathsf{T}}\mathbf{y} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H} = \mathbf{U}\mathbf{F}\mathbf{U}^{\mathsf{T}}$ and $h(\mathbf{x}_j) = \mathbf{H}_{jj}$.

4.1.4. *Relationship between leverage and the nullspace of* **X**

Recall that when we replace the validation sample $\mathbf{z}$ with the jth calibration sample in Equation (14), leverage simply becomes the sum of the squared score values:

$$h(\mathbf{x}_j) = ||\boldsymbol{\beta}||_2^2 = ||\mathbf{u}_{.,j}^{\mathrm{T}}||_2^2 = \sum_{i=1}^{r} u_{ji}^2$$

When the validation sample $\mathbf{z}$ does not coincide with a calibration sample, then $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_r]^{\mathrm{T}}$ can be regarded as *validation score* values. It is important to note that the validation score values ignore the nullspace portion $\mathbf{z}_2$ of $\mathbf{z}$ since

$$\boldsymbol{\beta} = \mathbf{S}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{z} = \mathbf{S}^{-1}\mathbf{V}^{\mathrm{T}}(\mathbf{z}_1 + \mathbf{z}_2) = \mathbf{S}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{z}_1$$

That is, the lengths of the minor axes associated with the ellipsoidal distribution of the data in the direction of the basis vectors of $\mathbf{V}_{\mathrm{null}}$ are zero. Hence, leverage only determines whether $\mathbf{z}_1$ is an outlier, not whether $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2$ is an outlier, i.e. leverage only identifies extreme samples that are far from the spectral space spanned by the calibration samples. If meaningful predictions are to be made, then the sparse calibration space inhabited by outliers must be populated by additional calibration samples. This aspect of leverage is both its strength and weakness as a diagnostic merit for outlier detection. On one hand, the covariance structure of the calibration data is rigorously adhered to. On the other hand, the nullspace component $\mathbf{z}_2$ could contribute significantly to the overall size of $||\mathbf{z}||_2$ but the leverage fit of this nullspace component will always be zero. The diagnostic merit known as the spectral *F*-ratio will pay special attention to this neglected orthogonal part of $\mathbf{z}$.

4.2. Spectral *F*-ratios

4.2.1. *Spectral residual via classical least squares*

The CLS-based spectral residual associated with a sample $\mathbf{z}$ is simply the nullspace component of $\mathbf{z}$ defined in Equation (12):

$$\mathbf{r}_{\mathrm{P}}(\mathbf{z}) = (\mathbf{I}_n - \mathbf{X}^{\dagger}\mathbf{X})\mathbf{z} = (\mathbf{I}_n - \mathbf{P})\mathbf{z} = \mathbf{z}_2 \tag{16}$$

where $\mathbf{r}_{\mathrm{P}}(\mathbf{z})$ denotes the spectral residual. The spectral residual provides useful diagnostic information: unexpected components in the validation spectrum that are not present in the calibration spectra tend to show up in the spectral residual.

4.2.2. *Spectral residual via regularized least squares*

When the pseudoinverse $\mathbf{X}^{\dagger}$ is replaced with the regularized inverse $\mathbf{X}^{\#}$ in Equation (16), we have

$$\mathbf{X}^{\#}\mathbf{X} = \mathbf{V}\mathbf{F}\mathbf{V}^{\mathrm{T}} = \mathbf{Q} \tag{17}$$

Here, $\mathbf{Q}$ maps the sample $\mathbf{z}$ onto a vector $\mathbf{q}_z \in \mathcal{S}_1$ such that

$$\mathbf{q}_z = \mathbf{Q}\mathbf{z} = \mathbf{V}\mathbf{F}\mathbf{V}^{\mathrm{T}}(\mathbf{z}_1 + \mathbf{z}_2) = \mathbf{V}\mathbf{F}\mathbf{V}^{\mathrm{T}}\mathbf{z}_1$$

Unlike $\mathbf{P} = \mathbf{V}\mathbf{V}^{\mathrm{T}}$, the matrix $\mathbf{Q} = \mathbf{V}\mathbf{F}\mathbf{V}^{\mathrm{T}}$ is not a projection matrix since $\mathbf{Q}^2 \neq \mathbf{Q}$ [15,21]. As a consequence, the spectral residual associated with $\mathbf{Q}$

$$\begin{aligned}\mathbf{r}_{\mathrm{Q}}(\mathbf{z}) &= (\mathbf{I}_n - \mathbf{Q})\mathbf{z} = \left(\mathbf{V}(\mathbf{I}_r - \mathbf{F})\mathbf{V}^{\mathrm{T}} + \mathbf{V}_{\mathrm{null}}\mathbf{V}_{\mathrm{null}}\right)(\mathbf{z}_1 + \mathbf{z}_2) \\ &= (\mathbf{z}_1 - \mathbf{q}_z) + \mathbf{z}_2\end{aligned} \tag{18}$$

differs from its counterpart $\mathbf{r}_{\mathrm{P}}(\mathbf{z})$ in Equation (16) in that $\mathbf{r}_{\mathrm{Q}}(\mathbf{z})$ is not strictly an element of $\mathcal{S}_2$. Instead, $\mathbf{r}_{\mathrm{Q}}(\mathbf{z})$ consists of two orthogonal components: the difference $\mathbf{z}_1 - \mathbf{q}_z \in \mathcal{S}_1$ and the orthogonal nullspace projection $\mathbf{z}_2 \in \mathcal{S}_2$. The key term in Equation (18) is $\mathbf{V}(\mathbf{I}_r - \mathbf{F})\mathbf{V}^{\mathrm{T}}$—contributions to the spectral residual come only from the minor axes of the hyperellipsoid and the nullspace projection $\mathbf{z}_2$ while contributions from the major axes are ignored. This stands in stark contrast to leverage where greater weight is given to leverage contributions associated with the major axes of the hyperellipsoid while contributions from the minor axes or the nullspace projection $\mathbf{z}_2$ are truncated or damped. Since the jth calibration sample $\mathbf{x}_j$ is already in $\mathcal{S}_1$, its spectral residual

$$\mathbf{r}_{\mathrm{Q}}(\mathbf{x}_j) = (\mathbf{I}_n - \mathbf{Q})\mathbf{x}_j = \left(\mathbf{V}(\mathbf{I}_r - \mathbf{F})\mathbf{V}^{\mathrm{T}} + \mathbf{V}_{\mathrm{null}}\mathbf{V}_{\mathrm{null}}\right)\mathbf{x}_j = (\mathbf{x}_j - \mathbf{q}_j) \in \mathcal{S}_1 \tag{19}$$

consists of one subspace component: the difference between $\mathbf{x}_j$ and $\mathbf{q}_j = \mathbf{Q}\mathbf{x}_j = \boldsymbol{VFS}\mathbf{u}_{.,j}^{\mathrm{T}}$.

4.2.3. *Spectral F-ratio via regularized least squares*

The diagnostic merit known as the spectral *F*-ratio is simply the ratio between the sum of the squared components of $\mathbf{r}_{\mathrm{Q}}(\mathbf{z})$ and the average of the sum of the squared components of $\mathbf{r}_{\mathrm{Q}}(\mathbf{x}_j)$ across all calibration samples. Using Equations (18) and (19), the spectral *F*-ratio can be mathematically expressed as

$$\delta(\mathbf{z}) = \frac{||\mathbf{r}_{\mathrm{Q}}(\mathbf{z})||_2^2}{\frac{1}{m}\sum_{j=1}^{m}||\mathbf{r}_{\mathrm{Q}}(\mathbf{x}_j)||_2^2} \tag{20}$$

where $\delta(\mathbf{z})$ represents the spectral *F*-ratio of $\mathbf{z}$. See the ASTM document in Reference [5] for details on how to use the spectral *F*-ratio in determining whether the corresponding sample is an outlier. Note that $||\mathbf{r}_{\mathrm{Q}}(\mathbf{x}_j)||_2^2$ is also equal to the jth diagonal element of the matrix $\mathbf{M}$ where

$$\mathbf{M} = \mathbf{X}(\mathbf{I}_n - \mathbf{Q})^2\mathbf{X}^{\mathrm{T}} = \mathbf{U}\mathbf{S}^2(\mathbf{I}_n - \mathbf{F})^2\mathbf{U}^{\mathrm{T}} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$$

and $\mathbf{D} = (\mathbf{I}_n - \mathbf{F})^2$. Since $\mathbf{U}$ is orthonormal, the trace of $\mathbf{M}$ is equal to the trace of $\mathbf{D}$ and the summation in the denominator of Equation (20) can be expressed as

$$\sum_{j=1}^{m}||\mathbf{r}_{\mathrm{Q}}(\mathbf{x}_j)||_2 = \mathrm{trace}(\mathbf{M}) = \mathrm{trace}(\mathbf{D}) = \sum_{i=1}^{r} s_i^2(1 - f_i)^2 \tag{21}$$

Using Equations (17) and (21), Equation (20) can now be compactly expressed via filter factors as

$$\delta(\mathbf{z}) = \frac{||(\mathbf{I}_n - \mathbf{V}\mathbf{F}\mathbf{V}^{\mathrm{T}})\mathbf{z}||_2^2}{\frac{1}{m}\sum_{i=1}^{r} s_i^2(1 - f_i)^2} \tag{22}$$

5. NUMERICAL RESULTS

5.1. Data set: corn

The data analyzed consist of the corn data set of Reference [39]. This corn data set consists of 80 samples of corn that were measured from 1100 to 2498 nm at 2 nm intervals on three (NIR) spectrometers designated as m5, mp5 and mp6. Note that every other wavelength was used to reduce the total number of wavelengths from 700 to 350. Reference values are provided for oil, protein, starch and moisture content. Protein is the prediction property studied in this paper and the spectra measured on instrument m5 serve as the primary calibration set.

5.2. Implementation

5.2.1. Software

MATLAB 7.04 (Release 2006a) was used in all numerical experiments [40]. In the interest of reproducible research, the code and data are available from the corresponding author upon request.

5.2.2. PLS algorithms

We now want to carefully distinguish how the PLS algorithms (filter factor and Krylov subspace versions) are implemented. In the chemometric literature, there are two types of PLS frameworks: *bidiagonalization* and *conventional*. These PLS frameworks are discussed in detail in Reference [33]. We use the bidiagonalization framework and, in particular, the LSQR algorithm of Paige and Saunders which performs k steps of the Lanczos bidiagonalization algorithm to solve the problem $\min ||\mathbf{Xb} - \mathbf{y}||_2$ [41]. For the filter factor based implementation of PLS, the MATLAB script `lsqr.m` (which implements the LSQR algorithm of Paige and Saunders) from *Regularization Tools* by Per Christian Hansen was used [42]. If one supplies `lsqr.m` with the singular values $s_1, s_2, \ldots, s_r$ of the calibration data $\mathbf{X}$, then the set of filter factors associated with the SVD expansion in Equations (6) and (8) is also returned as the matrix $\mathbf{F}_{\text{mat}} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k]$ such that $\mathbf{b}_{\text{PLS}} = \mathbf{VFS}^{-1}\mathbf{U}^T\mathbf{y}$ is the filter factor based solution for PLS solution where $\mathbf{F} = \text{diag}(\mathbf{f}_i)$. The filter factors in $\mathbf{F}$ are then used in tandem with the matrices $\mathbf{U}$, $\mathbf{S}$ and $\mathbf{V}$ (obtained from the SVD of $\mathbf{X}$) in order to compute the diagnostic measures. Note that the vector of regression coefficients $\mathbf{b}$ returned by `lsqr.m` and the regression coefficients returned by the SVD-based solution $\mathbf{VFS}^{-1}\mathbf{U}^T\mathbf{y}$ are the same. We will refer to this implementation of PLS as F-PLS to emphasize on the dependence on the filter factors. For the bidiagonalization approach using LSQR, the matrices $\tilde{\mathbf{U}}_{k+1}$, $\tilde{\mathbf{V}}_k$, $\tilde{\mathbf{R}}_k$ obtained from the Lanczos algorithm in Equation (10) will be used for both regression and computing leverages and spectral F-ratios. We will refer to this implementation of PLS as B-PLS to emphasize on the bidiagonal nature of the $\tilde{\mathbf{R}}_k$ matrix. In both F-PLS and B-PLS, re-orthonormalization using modified Gram–Schmidt orthogonalization was performed to maintain the orthogonal integrity of $\tilde{\mathbf{U}}_{k+1}$ and $\tilde{\mathbf{V}}_k$ [21].

5.2.3. Cross-validation procedure

Leverage and the spectral F-ratio can be calculated during the prediction phase as well as the calibration phase for diagnostic purposes. However, leverage and spectral F-ratio outliers in the calibration samples should be calculated in a cross-validated manner. The rationale is that when used for outlier detection, the outlier calibration samples left out in the cross validation will be poorly modeled and will consequently have large leverage and spectral F-ratio values. All n_{data} samples in the data set will be treated as calibration samples where n_{data} is the number of samples in the entire data set and n_{fold}-fold cross validation will then be performed where n_{fold} is the number of cross-validation folds. For example, the corn data set contains $n_{\text{data}} = 80$ samples and if $n_{\text{fold}} = n_{\text{data}}$, then leave-one-sample-out cross validation will be performed. In this paper, n_{fold} is set to 10. To avoid cross-validation results that are anecdotal to a particular ordering of samples, we will perform n_{order} rounds of n_{fold}-fold cross validation where $n_{\text{order}} = 100$ (with $n_{\text{fold}} = 10$). The fold membership of the samples will be randomly shuffled each round. We will now explain how the results across the cross-validation rounds will be combined.

The samples in each data set will be split into two parts: $(\mathbf{X}_{(j)}^{(i)}, \mathbf{y}_{(j)}^{(i)})$ and $(\mathbf{X}_{(j)}^{(-i)}, \mathbf{y}_{(j)}^{(-i)})$. The spectra $\mathbf{X}_{(j)}^{(i)}$ and response variables $\mathbf{y}_{(j)}^{(i)}$ correspond to the samples associated with the withheld ith cross-validation fold and the jth sample re-ordering while $\mathbf{X}_{(j)}^{(-i)}$ and $\mathbf{y}_{(j)}^{(-i)}$ correspond to the spectra and response variables, respectively, associated with the remaining samples. Note the abuse of notation for i and j. Previously i and j were used for the ıth principal component (or Krylov subspace) and jth sample in the data set. It will be clear from the context whether we mean principal component or sample, or cross-validation fold or sample re-ordering. The number of samples associated with $(\mathbf{X}_{(j)}^{(i)}, \mathbf{y}_{(j)}^{(i)})$ and $(\mathbf{X}_{(j)}^{(-i)}, \mathbf{y}_{(j)}^{(-i)})$ are m_i and $n_{\text{data}} - m_i$, respectively. In keeping with the notation of previous sections, the calibration data are set to $\mathbf{X} := \mathbf{X}_{(j)}^{(-i)}$ and $\mathbf{y} := \mathbf{y}_{(j)}^{(-i)}$ where $m = n_{\text{data}} - m_i$ and the validation spectrum $\mathbf{z}$ is one the m_i samples from $\mathbf{X}_{(j)}^{(i)}$. As described in Equation (1), both $\mathbf{X}_{(j)}^{(-i)}$ and $\mathbf{X}_{(j)}^{(i)}$ are mean-centered using the mean spectrum of the $\mathbf{X}_{(j)}^{(-i)}$ while $\mathbf{y}_{(j)}^{(-i)}$ is mean-centered with respect to its own mean response value. The calibration phase then consists of solving the linear system $\mathbf{X}_{(j)}^{(-i)}\mathbf{b}_{(j,k)}^{(i)} = \mathbf{y}_{(j)}^{(-i)}$ where $\mathbf{b}_{(j,k)}^{(i)}$ is the n-dimensional regression vector obtained using the kth regularization parameter (for PCR, using the first k principal components; for STR, using λ_k; and for F-PLS or B-PLS, using the first k dimensions of the Krylov subspace). Since there are n_{fold} cross-validation folds and n_{order} sample re-orderings, $n_{\text{fold}}n_{\text{order}}$ separate calibrations will be required for each MC method. Note that an SVD computation is performed per calibration.

The prediction phase results in the following m_i-dimension vector of estimations $\hat{\mathbf{y}}_{(j,k)}^{(i)} = \mathbf{X}_{(j)}^{(i)}\mathbf{b}_{(j,k)}^{(i)} + \bar{y}_{(j)}^{(-i)}\mathbf{1}_{m_i}$ where $\bar{y}_{(j)}^{(-i)}$ is the mean of the vector components in $\mathbf{y}_{(j)}^{(-i)}$. Collecting these estimated predictions, we can compute the root mean square error of cross validation (RMSECV) for the jth sample re-ordering and kth regularization parameter such that

$$e_{(j,k)} = \sqrt{\frac{1}{n_{\text{data}}}\sum_{i=1}^{n_{\text{fold}}} ||\hat{\mathbf{y}}_{(j,k)}^{(i)} - \mathbf{y}_{(j)}^{(i)}||_2^2}, \quad j = 1, \ldots, n_{\text{order}}, \quad k = 1, \ldots, q \tag{23}$$

where q is the number of regularization parameters. Since n_{fold} divides evenly into n_{data} for the corn set, q is equal to the rank of the calibration data $\mathbf{X}_{(j)}^{(-i)}$—which is the same across all folds and sample re-orderings. For the corn data set, $q = \min(m_i - 1, n) = \min(71, n) = 71$. For the regression vectors, the root mean square norm of cross validation (RMSNCV) will also be computed in a similar manner:

$$n_{(j,k)} = \sqrt{\frac{1}{n_{\text{fold}}}\sum_{i=1}^{n_{\text{fold}}} ||\mathbf{b}_{(j,k)}^{(i)}||_2^2}, \quad j = 1, \ldots, n_{\text{order}}, \quad k = 1, \ldots, q. \tag{24}$$

The median RMSECV and RMSNCV values for the kth regularization parameter, denoted as $e_{(k)}$ and $n_{(k)}$, respectively, will be calculated as

$$\begin{aligned} e_{(k)} &= \text{median}\{e_{(1,k)}, \ldots, e_{(n_{\text{order}},k)}\}, \\ n_{(k)} &= \text{median}\{n_{(1,k)}, \ldots, n_{(n_{\text{order}},k)}\}, \quad k = 1, \ldots, q \end{aligned} \tag{25}$$

For the projection methods (PCR, F-PLS and B-PLS), the number of regularization parameters q used depends on the rank of the calibration data. On the other hand, for STR, the number of λ values that one can use is arbitrary. In this case, the number of λ values is set to $q = 100$ where the maximum λ value is set to the largest singular value of $\mathbf{X}_{(j)}^{(-i)}$ (i.e. $\lambda_1 = s_1$) and the remaining λ values are chosen in an exponentially decaying fashion such that $\lambda_1 < \lambda_2 < \cdots < \lambda_q$. In both the projection and non-projection cases, the regularization parameter (k or λ_k) ranges from the most to the least amount of filtering/smoothing as k increases from 1 to q.

When calculating the leverages, $\mathbf{h}_{(j,k)}^{(i)}$ represents an m_i-dimensional vector of leverage values associated with the withheld samples for the kth regularization parameter, ith cross-validation fold and jth data re-ordering. The vector $\boldsymbol{\delta}_{(j,k)}^{(i)}$ denotes the spectral F-ratios in an analogous manner. These vectors are then concatenated to form n_{data}-dimensional vectors such that

$$\mathbf{h}_{(j,k)} = \begin{bmatrix} \mathbf{h}_{(j,k)}^{(1)} \\ \vdots \\ \mathbf{h}_{(j,k)}^{(n_{\text{fold}})} \end{bmatrix}, \quad \boldsymbol{\delta}_{(j,k)} = \begin{bmatrix} \boldsymbol{\delta}_{(j,k)}^{(1)} \\ \vdots \\ \boldsymbol{\delta}_{(j,k)}^{(n_{\text{fold}})} \end{bmatrix}, \quad j = 1, \ldots, n_{\text{order}}, \; k = 1, \ldots, q \tag{26}$$

The median leverage and spectral F-ratio values across the sample re-orderings for each regularization parameter, denoted as $\mathbf{h}_{(k)}$ and $\boldsymbol{\delta}_{(k)}$, respectively, are also n_{data}-dimensional vectors and are computed component-wise as $\mathbf{h}_{(k)} = \text{median}\{\mathbf{h}_{(1,k)}, \ldots, \mathbf{h}_{(n_{\text{order}},k)}\}$ and $\boldsymbol{\delta}_{(k)} = \text{median}\{\boldsymbol{\delta}_{(1,k)}, \ldots, \boldsymbol{\delta}_{(n_{\text{order}},k)}\}$, $k = 1, \ldots, q$. The median is used (as opposed to the mean) since some withheld samples for a specific data re-ordering have large leverage and spectral F-ratio values that are not consistent with the vast majority of the data reorderings. For consistency, the median measure was also used in Equation (25).

5.2.4. Regularization parameter selection

The diagnostic measures $\mathbf{h}_{(k)}$ and $\boldsymbol{\delta}_{(k)}$ are computed for each regularization parameter (k or λ_k). However, one generally commits to a particular or 'optimal' regularization parameter, say, for subsequent prediction on a validation data set. There are many criteria to select regularization parameters [17,43,44]. The simplest one finds the parameter that corresponds to the index k that minimizes the RMSECV values or

$$\mathbf{h}_{\text{opt}} = \mathbf{h}_{(k)}, \quad \boldsymbol{\delta}_{\text{opt}} = \boldsymbol{\delta}_{(k)}, \quad k = \arg\min_{1 \le i \le q} e_{(i)} \tag{27}$$

While this is not necessarily the best selection method, it represents a consistent method for comparison purposes. Unfortunately, the minimum RMSECV value on the curve $(k, e_{(k)})$ often occurs in a flat region. Hence, $e_{(k-k_0)} \approx e_{(k)}$ (for a small positive integer k_0) with $e_{(k-k_0)} > e_{(k)}$. This is not an uncommon scenario and will result in *under-regularization* and over-fitting. To avoid this pitfall, we will choose the first index k such that $e_{(k)}$ is beneath the threshold $a + c(b - a)$ where $a = \min\{e_1, \ldots, e_q\}$, $b = \max\{e_1, \ldots, e_q\}$ and $c = 0.05$. The constant c ($0 \le c << 1$) ensures that we are on the side of over-regularization as opposed to under-regularization. If $c = 0$, then we obtain the index given by Equation (27). We will refer to this regularization parameter selection strategy as the '$\min^+$' selection method.

5.3. Results

5.3.1. Model selections

Figures 1–3 show results for the corn data set. Figure 1 shows two ways to display the median RMSECV values. In the top subplot, the logarithm of the median RMSNCV and RMSECV coordinates, i.e. $(\log(n_{(k)}), \log(e_{(k)}))$, $k = 1, \ldots, q$, are plotted for each MC method and regularization parameter. This subplot illustrates the trade-off between the size of the regression vector (the amount of regularization) and the size of the prediction error, and is related to the L-curve which is used for regularization parameter selection [17,43]. The second way to display the median RMSECV values is the more commonplace one and is shown in the bottom subplot: RMSECV versus index k or the coordinates $(k, e_{(k)})$. For example, in the case of B-PLS, the coordinate $(k, e_{(k)})$ would correspond to the RMSECV associated with PLS factor k or the first k Krylov subspace dimensions. Qualitatively, all of the MC methods behave similarly in the top subplot. The plots for F-PLS and B-PLS are identical since they both compute the same regression coefficients. The regularization parameters corresponding to $\min^+$ selection method are $k = 14$ for PCR, $\lambda = \lambda_{31} = 0.0031$ for STR and $k = 8$ for both F-PLS and B-PLS. The coordinates associated with the indices 14, 31 and 8 have been enlarged on the curves in Figure 1. When RMSECV is *not* plotted on a *log*-scale, then the constant $c = 0.05$ seems reasonable. Note that for this STR-based calibration, the number of λ values was set to 100 such that the largest and smallest values were the largest singular value (s_1) and 0, respectively. The 98 intermediate λ values decreased from s_1 to 0 in an exponentially decaying manner. For the remaining MC methods, the number of principal components or latent factors was set to $q = \min(m_\theta - 1, n) = 71$. Although the regularization indices are not comparable to one another (e.g. PLS 'converges to an optimal' solution using fewer factors (the first eight Krylov subspace dimensions) than PCR (the first 14 principal components)), the solution norms $||\mathbf{b}||_2$ shown in the top subplot of Figure 1 are comparable and are in very close agreement across the methods.

5.3.2. Leverages and spectral F-ratios at selected models

Figure 2 shows the *normalized* leverage and spectral F-ratio outlier values, respectively, that were calculated according to the regularization parameters corresponding to the $\min^+$ selection method. By 'normalized', we mean that the outlier values were scaled to unit length, i.e. $\mathbf{h}_{\text{opt}}/||\mathbf{h}_{\text{opt}}||$ and $\boldsymbol{\delta}_{\text{opt}}/||\boldsymbol{\delta}_{\text{opt}}||$. In this way, we can facilitate a commensurable comparison of the outlier values across MC methods. For leverage, a great deal of consistency is achieved for all of the samples across the MC methods with samples 46, 47, 54, 55, 68, 72, 75 and 77 having the largest

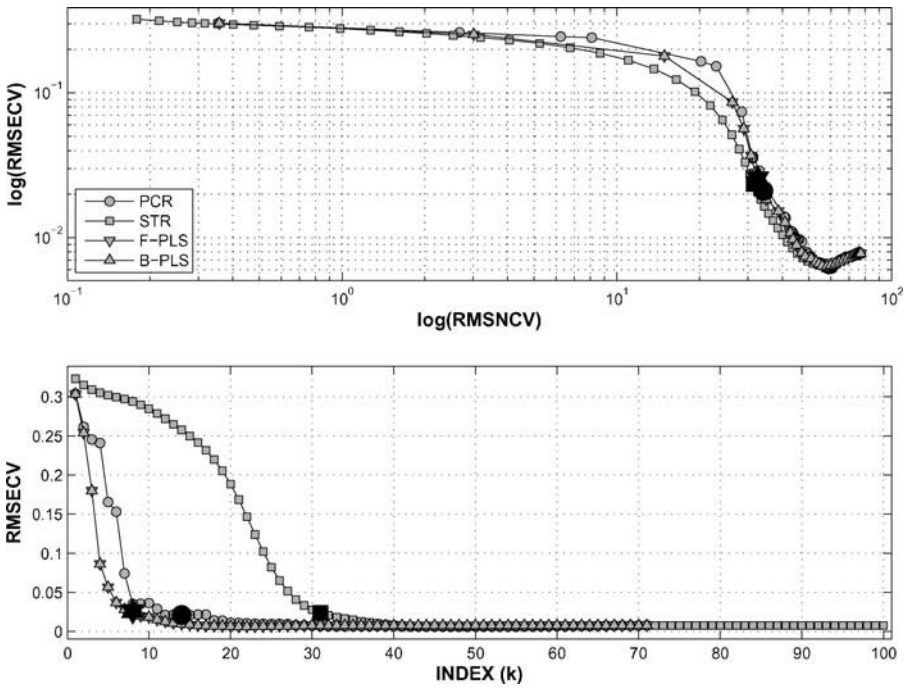

Figure 1. The top subplot shows the RMSNCV $n_{(k)}$ vs. the RMSECV $e_{(k)}$ across the four MC methods on a log–log plot while the bottom subplot shows the RMSECV $e_{(k)}$ versus the index k. The coordinates associated with the regularization parameters chosen by the $\min^{+}$ selection method are enlarged and black in color.

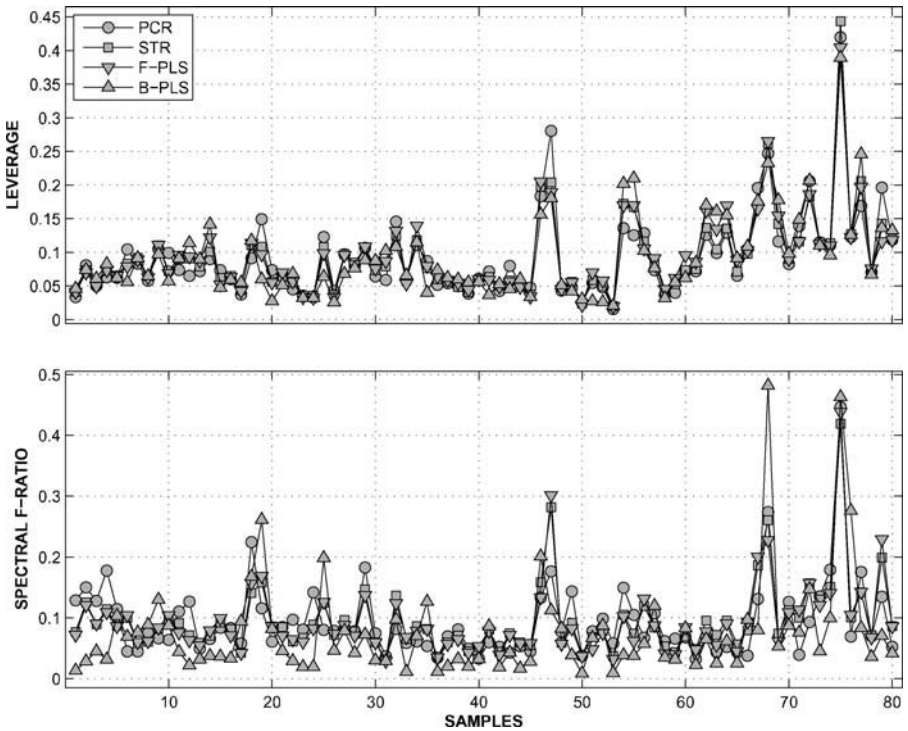

Figure 2. Sample vs. diagnostic measures across the four MC methods. The top and bottom subplots correspond to leverages ($\mathbf{h}_{opt}$) and spectral F-ratios (δ_{opt}), respectively.

leverage values. For the spectral F-ratios, the consistency was not as great as it was for leverage. For example, samples 46, 47, 68 and 75 had large spectral F-ratio values but there was no unanimous agreement on these samples being large across all MC methods.

On average, PCR, STR and F-PLS behave similarly with respect to the diagnostic merits in Figure 2. As a consequence, a natural question arises: How similar or dissimilar are the corresponding filter factors? To answer to this question, a single regression was performed using PCR, STR and F-PLS on *all* of the samples in the

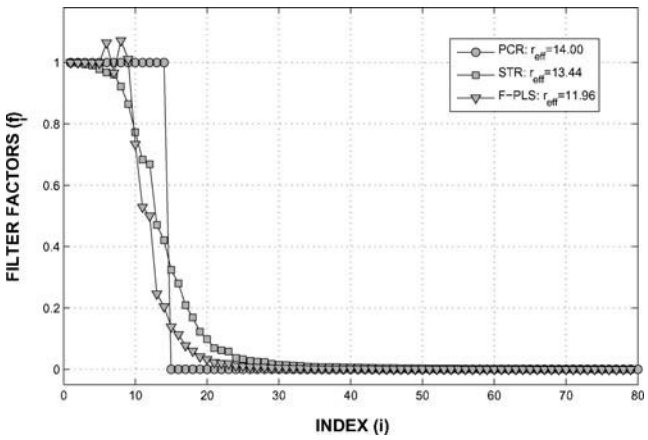

Figure 3. Index (i) vs. filter factors (f_i) using the regularization parameter obtained that minimizes RMSECV in Figure 1.

data set using the regularization parameters obtained from the min$^+$ selection method ($k = 14$ for PCR, $\lambda = \lambda_{31} = 0.0031$ and $k = 8$ the PLS-based methods). The resulting set of filter factor is shown in Figure 3. PCR exhibits a binary 0 or 1 filter factor representation while the filter factors for STR decay monotonically between 0 and 1. The filter factors for F-PLS exhibit a polynomial-like oscillation for $5 \leq k \leq 10$ decay, exceeding 1 in value, and then decaying in a manner similar to STR for $k > 10$. This type of behavior for F-PLS is typical [19,43]. As measured by the effective numerical rank (or the sum of the filter factors), PLS is the most parsimonious in terms of model complexity.

5.3.3. *Computational advantages*

If one is interested in regression only and not diagnostic measures, then Krylov subspace methods such as PLS are the preferred choice for large data sets since one does not have to explicitly form the subspace matrices $\tilde{\mathbf{U}}_{k+1}$, $\tilde{\mathbf{R}}_k$ and $\tilde{\mathbf{V}}_k$. This is not the case for the other MC methods since the SVD of the calibration data is required. (Note that other approaches exist for PCA that do not require the SVD such as NIPALS [45,46].) For large data sets, the SVD is computationally prohibitive. However, if one desires leverages and spectral F-ratios for a particular calibration model, then the subspace matrices are required. For a timing comparison, the `tic` and `toc` functions in MATLAB will be used to compute elapsed time. In particular, the timing will be based upon the median elapsed time for calibration (i.e. forming a regression vector and the subspace matrices) and the median elapsed time for leverage and spectral F-ratio calculations for a single calibration session on a cross-validation fold. For example, the median elapsed time for calibration is computed from the elapsed times across the $n_{\text{order}} n_{\text{fold}} = (100)(10) = 1000$ calibrations performed.

We will compare two MC methods—STR (100 regularization parameters) and B-PLS (80 factors)—on the corn data set. For STR, the median calibration time was 0.0100 s (0.0087 s of which was calculating the SVD) while the median time for calculating leverages and spectral F-ratios was 0.0190 s. For B-PLS, the median calibration time was 0.0258 s while it took a median time of 0.1421 s to calculate the leverages and spectral F-ratios. The calibration time for PLS was considerably slowed down due to the re-orthonormalization requirement we imposed on the subspace matrices $\tilde{\mathbf{U}}_{k+1}$ and $\tilde{\mathbf{V}}_k$. Note that we would expect B-PLS to outperform (time-wise) SVD-based MC methods for substantially larger data sets. Calculating leverages and spectral F-ratios via the filter factor approach results in a speed-up factor of approximately $0.1421/0.0190 \approx 7.5$. This speed-up is achieved in two ways. First, there is no inversion required of a bidiagonal matrix for STR. Second, the calculation of the mean spectral residual for the calibration spectra is a simple sum involving only singular value and filter factors—see Equation (22). Although the unit of time considered is a single calibration session on a cross-validation fold, the cumulative efficiency of the filter factor approach across many rounds of cross validation can be considerable.

6. CONCLUSION AND FUTURE WORK

We propose a new algorithm that allows one to compute the MC diagnostic measures, leverages and spectral F-ratios, using a filter factor representation. This framework extends these diagnostic measures from projection-based based methods (PCR and PLS) to any regression method that admits a filter factor representation—projection-based or not. In addition to being extensible, the filter factor representation is also computationally thrifty. The MC methods outlined here are simple regularization schemes based upon the SVD. However, SVD does not scale well to large data sets. Recently discussed randomized versions of SVD are considerably more efficient and reliable than the classical/deterministic SVD version and are also amenable to parallelization [47,48]. Being probabilistic, these schemes have a finite probability of failure but, in most cases, this probability is negligible (e.g. 10^{-17}). Other future work would abandon the SVD altogether in favor of 'non-standard' factorizations that may be more appropriate when scaling up to massive data sets, incorporating spectroscopic *a priori* information, updating/downdating rows (samples) or columns (wavelengths) or performing wavelength selection.

REFERENCES

1. De Maesschalck R, Estienne F, Verdú-Andrés J, Candolfi A, Centner V, Despagne F, Jouan-Rimbaud D, Walczak B, Massart D, de Jong S, de Noord O, Puel C, Vandeginste B. The development of calibration models for spectroscopic data using principal component regression. Vrije Universiteit Brussel, Department of Analytical Chemistry and Pharmaceutical Technology. http://www.vub.ac.be/fabi/tutorials.html
2. Geladi P, Kowalski B. Partial least squares regression: a tutorial. *Anal. Chim. Acta* 1986; **185**: 1–17.
3. Tikhonov A. Solution of incorrectly formulated problems and the regularization method. English translation of *Dokl. Akad. Nauk.* 1963; **151**: 501–504.
4. Hoerl A. Application of ridge analysis to regression problems. *Chem. Eng. Prog.* 1962; **58**: 54–59.
5. Standard Practice for Validation of the Performance of Multivariate Process Infrared Spectrophotometers. *ASTM International, Designation D6122-06e1* 2006.
6. Naes T, Isaksson T, Fearn T, Davies T. *A User-friendly Guide to Multivariate Calibration and Classification*. NIR Publications: Chichester, UK, 2002.
7. Pell R. Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemom. Intell. Lab. Syst.* 2000; **52**(1): 87–104.
8. Daszykowski M, Kaczmarek K, Heyden V, Walczak B. Robust statistics in data analysis: A review. Basic concepts. *Chemom. Intell. Lab. Syst.* 2007; **85**(2): 203–219.
9. Ferré J. *Regression Diagnostics in Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Vol. 3, Brown S, Tauler R, Walczak B (eds). Elsevier: The Netherlands, 2009.
10. Rousseeuw P, Leroy A. *Robust Regression and Outlier Detection*. Wiley: New York, NY, 1987.

11. *Theory and Applications of Recent Robust Methods*, Hubert M, Pison G, Struyf A, Van Aelst S (eds). Series: Statistics for Industry and Technology. Birkhäuser Verlag: Basel, Switzerland, 2004.
12. Verboven S, Hubert M. LIBRA: a MATLAB Library for Robust Analysis. *Chemom. Intell. Lab. Syst.* 2005; **75**: 127–136.
13. Aggarwal C, Yu P. Outlier detection for high dimensional data. *ACM SIGMOD Conference*, 2001.
14. Hadi A, Imon AR, Werner M. Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics* 2009; **1**(1): 57–70.
15. Demmel J. *Applied Numerical Linear Algebra*. SIAM Press: Philadelphia, PA, 1997.
16. Golub G, Kahan W. Calculating the singular values and pseudo-inverse of a matrix. *SIAM J. Numer. Anal., Ser. B* 1965; **2**: 205–224.
17. Hansen PC. *Rank-deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion*. SIAM Press: Philadelphia, PA, 1998.
18. Helland I. On the structure of Partial Least Squares. *Commun. Stat. Simul. Comput.* 1988; **17**: 581–607.
19. Lingjærde O, Christopherson N. Shrinkage structure of partial least squares. *Scand. J. Stat.* 2000; **27**: 459–473.
20. Manne R. Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemom. Intell. Lab. Syst.* 1987; **2**: 187–197.
21. Trefethen LN, Bau III D. *Numerical Linear Algebra*. SIAM Press: Philadelphia, PA, 1997.
22. Walker E, Birch J. Influence measures in ridge regression. *Technometrics* 1988; **30**(2): 221–227.
23. Steece B. Regressor space outliers in ridge regression. *Commun. Stat.—Theory Methods* 1986; **15**(12): 3599–3605.
24. Shi L, Wang X. Local influence in ridge regression. *Comput. Stat. Data Anal.* 1999; **31**: 341–353.
25. Andries E, Hagstrom T, Atlas S, Willman C. Regularization strategies for hyperplane classifiers: application to cancer classification. *J. Bioinform. Comput. Biol.* 2007; **5**(1): 79–104.
26. Hastie T, Tibsharani R, Friedman J. *The Elements of Statistical Learning*. Springer-Verlag: New York, NY, 2001.
27. Kalivas JH. Interrelationships of multivariate regression methods using eigenvector basis sets. *J. Chemom.* 1999; **13**: 111–132.
28. Kalivas JH. Basis sets for multivariate calibration. *Anal. Chim. Acta* 2001; **428**: 31–40.
29. Tibshirani R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 1996; **58**(1): 267–288.
30. Lawson CL, Hanson RJ. *Solving Least Squares Problems*. Prentice Hall Press: Englewood Cliffs, NJ, 1974.
31. DiFoggio R. Desensitizing models using covariance matrix transforms or counter-balanced distortions. *J. Chemom.* 2005; **19**(4): 203–215.
32. DiFoggio R. Influencing models to improve their predictions of standard samples. *J. Chemom.* 2007; **21**(5–6): 208–214.
33. Pell RJ, Ramos LS, Manne R. The model space of partial least squares regression. *J. Chemom.* 2007; **21**(3–4): 165–172.
34. van der Vorst H. *Iterative Krylov Methods for Large Linear Systems*. Cambridge University Press: Cambridge, UK 2003.
35. Wentzell P, Vega-Montoto L. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemom. Intell. Lab. Syst.* 2003; **65**: 257–279.
36. Kelley CT. *Iterative Methods for Linear and Nonlinear Equations*. SIAM Press: Philadelphia, PA, 1995.
37. Faber N, Ferré J. On the numerical stability of two widely used PLS algorithms. *J. Chemom.* 2008; **22**(2): 101–105.
38. Stewart GW. *Matrix Algorithms: Basic Decompositions*. SIAM Press: Philadelphia, PA, 1998.
39. BM Wise B., Gallagher NB, Bro R, Shaver JM. *PLS_Toolbox 3.0 for Use with MATLAB*. Eigenvector Research: Manson, WA, 2003.
40. MATLAB. *User's Guide*. The MathWorks, Inc.: Natick, MA 01760, 1992.
41. Paige C, Saunders M. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Softw.* 1982; **8**: 43–71.
42. Hansen PC. Regularization tools: a Matlab package for analysis and solution of discrete Ill-posed problems. *Numer. Algorithms* 1994; **6**: 1–35.
43. Kalivas JH. Calibration methodologies. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Vol. 3, Brown S, Tauler R, Walczak B (eds). Elsevier: The Netherlands, 2009.
44. Golub G, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 1979; **21**: 215–223.
45. Wold H. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, Krishnaiah P (ed.). Academic Press: New York, NY, 1966.
46. Wold H. Path models with latent variables: The NIPALS approach. In *Quantitative Sociology: International perspectives on mathematical and statistical model building*, Blalock H, Aganbegian A, Borodkin F, Boudon R, Cappecchi V (eds). Academic Press: NY, 1975.
47. Liberty E, Woolfe F, Martinsson P-G, Rokhlin V, Tygert M. Randomized algorithms for the low-rank approximation of matrices. *Proc. Natl Acad. Sci. U.S.A.* 2007; **104**(51): 20167-–20172.
48. Rokhlin V, Szlam A, Tygert M. A randomized algorithm for principal component analysis. *SIAM J. Matrix Anal. Appl.* 2009; **31**(3): 1100—1124.
49. Jackson J. *A User's Guide to Principal Components*. Wiley: New York, NY, 1991.
50. Hestenes M, Stiefel E. Methods of conjugate gradient for solving linear systems. *J. Res. Nat Bur. Stand.* 1952; **49**: 409–436.
51. van der Sluis A, van der Vorst H. The rate of convergence of conjugate gradients. *Numer. Math.* 1986; **48**: 543–560.

APPENDIX A: DERIVATION OF THE FILTER FACTORS FOR PCR

For PCR, the subspace $\mathcal{S}$ in Equation (4) consists of the space spanned by the first k right singular vectors of $\mathbf{X}$, i.e. $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$. The vectors $\mathbf{c}_i = \mathbf{X}\mathbf{v}_i$ are called the principal components (PCs) of $\mathbf{X}$. The direction $\mathbf{c}_1 = \mathbf{X}v_1$ contains the largest sample variance amongst all normalized linear combinations of $\mathbf{X}$ and all subsequent PCs have maximum variance subject to being orthogonal to previous PCs [49]. If $\mathbf{X}_k = \sum_{i=1}^{k} s_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$, then PCR amounts to the CLS minimization of $||\mathbf{X}_k\mathbf{b} - \mathbf{y}||_2^2$ with the solution being identical to Equation (2) except that the last $r - k$ terms have been truncated:

$$\mathbf{b}_{\mathrm{PCR}} = \mathbf{X}_k^{\dagger}\mathbf{y} = \sum_{i=1}^{k} \alpha_i \mathbf{v}_i = \mathbf{V}\mathbf{F}\boldsymbol{\alpha}, \tag{28}$$

$$\mathbf{F} = \mathrm{diag}(\mathbf{1}_k, \mathbf{0}), \quad f_i = \begin{cases} 1, & i = 1, \ldots, k \\ 0, & i = k+1, \ldots, r \end{cases} \tag{29}$$

PCR is an appropriate regularization strategy if one can reliably estimate the numerical rank k of the data such that there is a well-defined gap between s_k and s_{k+1}.

APPENDIX B: DERIVATION OF THE FILTER FACTORS FOR STR

If there is no well-determined gap in the singular value spectrum, truncation of the SVD expansion may not lead to the best-regularized solution. If we truncate too early then we may lose information, and if we include too many terms then the solution can become unstable in the presence of noise. A *softer* threshold approach based upon a set of filter factors that are not binary (0 or 1) may be required such that terms with small singular values are damped to a greater degree than terms with larger singular values. TR allows for such flexibility. In the STR version of TR, the solution of Equation (5) is obtained by setting the gradient of $\phi(\mathbf{b})$ equal to zero and solving for $\mathbf{b}$ using the relation $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$:

$$\mathbf{b}_{\mathrm{STR}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda^2\mathbf{I}_r)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{V}(\mathbf{S}^2 + \lambda^2\mathbf{I}_r)^{-1}\mathbf{S}\mathbf{U}^{\mathrm{T}}\mathbf{y} = \mathbf{V}\mathbf{F}\boldsymbol{\alpha} \tag{30}$$

$$\mathbf{F} = (\mathbf{S}^2 + \lambda^2\mathbf{I}_r)^{-1}\mathbf{S}^2, \quad f_i = \frac{s_i^2}{s_i^2 + \lambda^2}, \; i = 1, \ldots, r \tag{31}$$

Note that for any $\lambda \geq 0$ the filter factors are bounded between 0 and 1. When $\lambda = 0$, the STR solution reverts back to the CLS solution since $\mathbf{F} = \mathbf{I}_r$.

APPENDIX C: DERIVATION OF THE FILTER FACTORS FOR PLS

The construction of the filter factors for PLS is more complicated. PLS is a *Krylov subspace* algorithm, i.e. the solution to Equation (4) is restricted to lie in the Krylov subspace $\mathcal{S} = \mathcal{K}_k(\mathbf{G}, \mathbf{d})$ of dimension k and is defined as

$$\mathcal{K}_k(\mathbf{G}, \mathbf{d}) = \text{span}\{\mathbf{d}, \mathbf{G}\mathbf{d}, \ldots, \mathbf{G}^{k-1}\mathbf{d}\}$$

with $\mathbf{G}$ being a positive (semi)definite matrix [18]. Here, the matrix $\mathbf{G}$ and the vector $\mathbf{d}$ are associated with the CLS-based normal equations $\mathbf{G}\mathbf{b}_{\text{CLS}} = \mathbf{d}$ such that $\mathbf{G} = \mathbf{X}^{\text{T}}\mathbf{X}$ and $\mathbf{d} = \mathbf{X}^{\text{T}}\mathbf{y}$. The most well-known Krylov subspace algorithm is the conjugate gradient (CG) algorithm and it has come into widespread use for solving large-scale systems of linear equations [36,50]. In the numerical analysis, statistics and chemometrics literature, there are many ways to implement both PLS and CG. However, the commonality that these two algorithms share is that they are exactly the same numerical procedure when implemented via the Lanczos bidiagonalization algorithm [16–18,20,33]. We will take this approach and will use $\mathbf{b}_{\text{PLS}}$ to refer to the CG solution obtained via the LBP. We now briefly examine how the filter factors associated with PLS are constructed. The original filter factor derivation for CG (and by extension PLS) first appeared in Reference [51].

Since the CG solution vector $\mathbf{b}_{\text{PLS}}$ lies in the Krylov subspace $\mathcal{K}_k(\mathbf{G}, \mathbf{d})$, it can be rewritten as the following linear combination:

$$\mathbf{b}_{\text{PLS}} = \sum_{i=0}^{k-1} \gamma_i \mathbf{G}^i \mathbf{d} = \sum_{i=0}^{k-1} \gamma_i \mathbf{G}^i (\mathbf{G}\mathbf{b}_{\text{CLS}}) = \sum_{i=0}^{k-1} \gamma_i \mathbf{G}^{i+1} \mathbf{b}_{\text{CLS}}$$

for some coefficients γ_i. As a result, the error between the CLS solution and the CG solution vector can be expressed as a matrix polynomial

$$\mathbf{b}_{\text{CLS}} - \mathbf{b}_{\text{PLS}} = \mathbf{b}_{\text{CLS}} - \sum_{i=1}^{k-1} \gamma_i \mathbf{G}^{i+1} \mathbf{b}_{\text{CLS}} = p(\mathbf{G})\mathbf{b}_{\text{CLS}},$$

$$p(\theta) = 1 - \sum_{i=1}^{k-1} \gamma_i \theta^{i+1} \tag{32}$$

where the polynomial $p(\theta)$ has degree k and satisfies $p(0) = 1$. The filter factors associated with PLS ultimately derive their damping behavior from the oscillatory nature of a particular type of polynomial used in Equation (32).

Minimizing $||\mathbf{X}\mathbf{b} - \mathbf{y}||_2^2$ over any subset $\mathcal{S}$ in $\mathbb{R}^n$ is the same as minimizing $||\mathbf{b}_{\text{CLS}} - \mathbf{b}||_{\mathbf{G}}^2$ over $\mathcal{S}$ (where $||\mathbf{x}||_{\mathbf{G}}^2 = \mathbf{x}^{\text{T}}\mathbf{G}\mathbf{x}$ is known as the *induced* norm) since

$$\begin{aligned} \min_{b \in \mathcal{S}} ||\mathbf{X}\mathbf{b} - \mathbf{y}||_2^2 &= \min_{b \in \mathcal{S}} \mathbf{b}^{\text{T}}\mathbf{G}\mathbf{b} - 2\mathbf{d}^{\text{T}}\mathbf{b} + \mathbf{y}^{\text{T}}\mathbf{y} \\ &= \min_{b \in \mathcal{S}} \mathbf{b}^{\text{T}}\mathbf{G}\mathbf{b} - 2\mathbf{d}^{\text{T}}\mathbf{b} + \mathbf{b}_{\text{CLS}}^{\text{T}}\mathbf{G}\mathbf{b}_{\text{CLS}} \\ &= \min_{b \in \mathcal{S}} ||\mathbf{b} - \mathbf{b}_{\text{CLS}}||_{\mathbf{G}}^2 \end{aligned} \tag{33}$$

If $\mathcal{S} = \mathcal{K}_k(\mathbf{G}, \mathbf{d})$ and if we substitute Equation (32) into the right-hand side of Equation (33), then $\mathbf{b}_{\text{PLS}}$ can be expressed as the solution of the following polynomial approximation problem:

$$\min_{b \in \mathcal{S}} ||\mathbf{X}\mathbf{b} - \mathbf{y}||_2^2 = \min_{p \in P_k, p(0)=1} ||p(\mathbf{G})\mathbf{b}_{\text{CLS}}||_{\mathbf{G}} \tag{34}$$

where P_k denotes the set of polynomials of degree k and $p(0) = 1$. The polynomial which minimizes Equation (34) is known as the Ritz polynomial

$$\mathcal{R}_k(\theta) = \left(\frac{\theta_1^k - \theta}{\theta_1^k}\right) \cdots \left(\frac{\theta_k^k - \theta}{\theta_k^k}\right) = \prod_{i=1}^{k} \left(\frac{\theta_i^k - \theta}{\theta_i^k}\right)$$

where the *Ritz values* θ_i^k are the corresponding polynomial zeros [21,36,51]. The orthonormality of $\mathbf{V}$ allows us to rewrite any power of $\mathbf{G}$ as $\mathbf{G}^i = \mathbf{V}(\mathbf{S}^2)^i\mathbf{V}^{\text{T}}$. By extension, any matrix polynomial involving $\mathbf{G}$ could be expressed as $p(\mathbf{G}) = \mathbf{V}p(\mathbf{S}^2)\mathbf{V}^{\text{T}}$. If we solve for $\mathbf{b}_{\text{PLS}}$ in Equation (32) using the relation $\mathbf{b}_{\text{CLS}} = \mathbf{V}\boldsymbol{\alpha}$ in Equation (3) and $\mathcal{R}_k(\mathbf{G}) = \mathbf{V}\mathcal{R}_k(\mathbf{S}^2)\mathbf{V}^{\text{T}}$, then we finally arrive at the filter-factor-based solution for CG iteration number k (or more commonly known as the kth PLS factor):

$$\begin{aligned} \mathbf{b}_{\text{PLS}} &= \mathbf{b}_{\text{CLS}} - \mathcal{R}_k(\mathbf{G})\mathbf{b}_{\text{CLS}} = \left(\mathbf{I}_n - \mathbf{V}\mathcal{R}_k(\mathbf{S}^2)\mathbf{V}^{\text{T}}\right)\mathbf{b}_{\text{CLS}} \\ &= \mathbf{V}\left(\mathbf{I}_r - \mathcal{R}_k(\mathbf{S}^2)\right)\boldsymbol{\alpha} = \mathbf{V}\mathbf{F}\boldsymbol{\alpha} \end{aligned} \tag{35}$$

$$\mathbf{F} = \mathbf{I}_r - \mathcal{R}_k(\mathbf{S}^2), \quad f_i^k = 1 - \mathcal{R}_k(s_i^2) = 1 - \prod_{i=1}^{k} \left(\frac{\theta_i^k - s_i^2}{\theta_i^k}\right) \tag{36}$$

Unlike their PCR and STR counterparts, the PLS-based filter factors depend not only on $\mathbf{X}$ but also on the response variable $\mathbf{y}$ since $\mathbf{b} \in \mathcal{K}_k(\mathbf{G}, \mathbf{d}) = \mathcal{K}_k(\mathbf{X}^{\text{T}}\mathbf{X}, \mathbf{X}^{\text{T}}\mathbf{y})$. Moreover, the filter factors can take on values outside the interval [0, 1] since f_i^k is the value of a kth degree polynomial evaluated at $\theta = s_i^2$.