Research article

Journal of CHEMOMETRICS

Received: 14 July 2015, Revised: 3 January 2016, Accepted: 4 January 2016, Published online in Wiley Online Library: 17 February 2016

(wileyonlinelibrary.com) DOI: 10.1002/cem.2778

Using the L_1 norm to select basis set vectors for multivariate calibration and calibration updating

Parviz Shahbazikhah[a,b], John H. Kalivas[a]*, Erik Andries[c,e] and Trevor O'Loughlin[d]

With projection based calibration approaches, such as partial least squares (PLS) and principal component regression (PCR), the calibration space is spanned by respective basis vectors (latent vectors). Up to rank k basis vectors are formed where $k \leq \min(m,n)$ with m and n denoting the number of calibration samples and measured variables. The user needs to decide how many and which respective basis vectors (tuning parameters). To avoid the second issue, basis vectors are selected top-down starting with the first and sequentially adding until model criteria are satisfied. Ridge regression (RR) avoids the issues by using the full set of basis vectors. Another approach is to select a subset from the total available. The presented work develops a process based on the L_1 vector norm to select basis vectors. Specifically, the L_1 norm is used to select singular value decomposition (SVD) basis set vectors for PCR (LPCR). Because PCR, PLS, RR, and others can be expressed as linear combination of the SVD basis vectors, the focus is on selection and comparison using the SVD basis set. Results based on respective tuning parameter selections and weights applied to the SVD basis vectors for LPCR, top-down PCR, correlation PCR (CPCR), PLS, and RR are compared for calibration and calibration updating using spectroscopic data sets. The methods are found to predict equivalently. In particular, the L_1 norm produces similar results to those obtained by the well-studied CPCR process. Thus, the new method provides a different theoretical framework than CPCR for selecting basis vectors. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: sparse model; principal component selection; multivariate calibration; calibration maintenance; model updating

1. INTRODUCTION

In analytical chemistry, developing a mathematical relationship between chemical and/or physical variables (analytes) and measured spectra is common [1–3]. The typical linear mathematical relationship for multivariate calibration is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ denotes an $m \times 1$ vector of quantitative reference analyte information for m calibration samples, $\mathbf{X}$ symbolizes the $m \times n$ matrix of respective spectra measured over n wavelengths, $\mathbf{b}$ represents an $n \times 1$ model vector, and $\mathbf{e}$ signifies the $m \times 1$ vector of normally distributed errors with mean zero and covariance matrix $\sigma^2\mathbf{I}$ with $\mathbf{I}$ being the $m \times m$ identity matrix. Multivariate calibration seeks to estimate $\mathbf{b}$, by $\hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y}$ where $\mathbf{X}^{+}$ is a generalized inverse of $\mathbf{X}$.

Different methods are used to form a generalized inverse. Three approaches are partial least squares (PLS), principal component regression (PCR), and ridge regression (RR) [1–3]. The methods can be used in full wavelength mode (no wavelength selection is required), but other tuning parameters must be ascertained. Tuning parameters for PLS and PCR are the number of respective basis vectors. Other terms for basis vectors are latent vectors or variables (LVs) or principal components (PCs). The role of the tuning parameter value for PCR and PLS is to reduce the dimensionality of the calibration space, and as a result, shrink the regression vector relative to using the full space, i.e. the model vector 2-norm (L_2 norm) $\|\hat{\mathbf{b}}\|_2$ increases as the number of basis vectors included increases. For RR, the tuning parameter is the ridge value that also controls how much of the calibration space is used and results in a shrunken regression vector.

In all three methods, the tuning parameters adjust the model vector direction and magnitude balancing the underlying model selectivity/sensitivity tradeoff and hence, the bias/variance tradeoff

* Correspondence to: John H. Kalivas, Department of Chemistry, Idaho State University, Pocatello, ID 83209, USA.
E-mail: kalijohn@isu.edu

a P. Shahbazikhah, J. H. Kalivas
Department of Chemistry, Idaho State University, Pocatello, ID 83209, USA

b P. Shahbazikhah
Metrohm Canada, 4160 Sladeview Crescent, #6, Mississauga, ON L5L 0A1, Canada

c E. Andries
Center for Advanced Research Computing, University of New Mexico, Albuquerque, NM 87106, USA

d T. O'Loughlin
Department of Physics, Texas Tech University, Lubbock, TX 79409, USA

e E. Andries
Department of Mathematics, Central New Mexico Community College, Albuquerque, NM 87106, USA

[4]. Specifically, the inverse of the model vector L_2 norm is a measure of the model sensitivity [5,6] and is adjusted as the tuning parameters change with a tradeoff to the model vector selectivity. As the model vector model vector L_2 norm increases, the potential for greater prediction variance also increases while the prediction error for the calibration samples decreases (less bias).

The methods of PCR and PLS use different basis sets to span the calibration space. Up to rank k respective basis vectors are formed to span the complete calibration space including the noise where $k \le \min(m,n)$. Once the k respective basis vectors have been formed, the user needs to decide how many and which ones to form the final projection based model. To avoid the second issue for PLS and PCR, basis vectors are selected in a top-down manner starting with the first basis vector formed and sequentially adding more until acceptable values for measures of model quality are attained. For the method of PCR, this is sometimes referred to as top-down PCR.

For PCR, studies have been performed where a subset from the k PCR basis vectors, not necessarily in the top-down order, have been selected [7–15]. A well-studied approach selects the basis vectors based on the correlations to the prediction property and is termed correlation PCR (CPCR) [12–15].

A variant of Tikhonov regularization (TR) [16] accomplishes variable selection by using an L_1 norm penalty on the model vector ($\|\mathbf{b}\|_1$) instead of the L_2 norm as with RR [1,17–22]. This variable selection variant of TR is commonly referred to as least absolute shrinkage and selection operator (LASSO) [21]. Presented in this paper is a TR process using the L_1 to select PCR basis vectors best balancing the bias/variance tradeoff thereby potentially reducing prediction errors. The method is referred to as LASSO PCR (LPCR). Using several measures of model quality and spectroscopic data sets, LPCR is compared to PCR, CPCR, PLS, and RR. Subset selection with the L_1 norm should not be confused with other methods that have sparsified regression methods such as PCR and PLS [23–27]. In these situations, the intent is to create sparse basis vectors such that when used, a sparse regression vector is formed that is essentially variable (wavelength) selected. For the data sets studied in this paper, the intent is to be full wavelength based, but the sparseness is in which non-sequential PCR basis vectors are used.

The regression vectors for PCR, PLS, and RR can be expressed as linear combinations of the **V** singular vectors of **X** from the singular value decomposition (SVD) of **X** ($\mathbf{X} = \mathbf{U\Sigma V}^T$) [2,3,28–35]. These SVD basis vectors are also the PCR basis vectors. By using a common basis set to express model vectors, similarities and differences between the various methods can be observed. The respective model weight values for the common basis set indicate when a model vector will be shrunken or expanded relative to another model vector. Also evaluated in this paper are the SVD basis vector weights for the LPCR, PCR, CPCR, PLS, and RR model vectors.

A significant problem in multivariate calibration using spectroscopic data is the changing measurement conditions (secondary conditions) from when the calibration samples were measured (primary conditions) and the model was formed. Thus, methods have been and continue to be developed to update a model formulated in the original primary conditions to now predict new samples measured in the secondary conditions [36]. While other processes have been developed, model updating is the method studied in this paper. Specifically, a hybrid of LPCR for calibration with a TR model updating method [22]. Results are compared with CPCR, PCR, PLS, and RR as adapted for calibration updating.

2. THEORY

Because all the methods presented in this paper require tuning parameter selection, a brief discussion is first presented on the tuning parameter selection process for this study. This discussion is followed by descriptions of the new calibration processes.

2.1. Selecting final tuning parameter values and assessing effective rank (ER)

Numerous approaches exist to determine tuning parameter values, e.g. the number of respective basis vectors for PCR and PLS, or the ridge value for RR [37–46]. Typically, a single criterion is used based on some sort of cross-validation (CV) procedure such as leave-one-out CV (LOOCV) or leave multiple out CV (LMOCV) to compute the root mean square error of CV (RMSECV) values. Another approach implemented in this study incorporates graphical analysis of multiple model quality measures as a tuning parameter varies. Typically, a model quality measure of prediction error (such as RMSECV or the RMSE of calibration RMSEC) or model fit (such as R^2) are plotted against a measure of model complexity (such as the model vector magnitude measured by the L_2 vector norm or 2-norm) [44–46]. When RMSEC is used, the plotted curves are L-shaped, and the graphical analysis is termed the L-curve method [47,48]. The tuning parameters forming models with the better bias/variance tradeoff are in the corner region of the L-curve. Such L-curve related plots are evaluated in this paper and used to select tuning parameters for all the calibration processes.

Another model diagnostic measure to distinguish the degree of model complexity is the ER or effective degrees of freedom [8,48–54]. One computational approach for ER, termed the generalized degrees of freedom (GDF), uses a Monte Carlo method [54]. The GDF algorithm for the ER used in this paper first adds normally distributed noise (δ) to each sample in vector **y** N times, obtaining respective N vectors of $\hat{\mathbf{y}}$ from the tuning parameter specific models formed using the corresponding perturbed calibration **y**. Next the linear regression slope ω_i for the ith sample is obtained for the equation $\hat{y}_i = \rho + \omega_i \delta_{ij}$ for $j = 1,\ldots,N$ with intercept ρ. The ER for the particular tuning parameter is then computed by $\mathrm{ER} = \sum_{i=1}^{m} \omega_i$ for the m samples in **y**. While the method requires a value for δ_i, practical experiences shows that ER values are generally invariant to the actual value, i.e. a large range of values produces the same results. The ER provides another measure of model complexity for an impartial graphical comparison between calibration methods.

The statistical significance of basis vectors as it is entered into the model can be assessed using a randomization test [37] as well as other statistical testing processes [7,8,14]. These processes were not evaluated for any of the calibration methods. Instead, measures of model quality are tracked.

2.2. Singular vector basis set and model vector basis weights (β)

The approach of PCR is to estimate the regression vector **b** by only including the top d basis vectors pertinent to modeling the analyte from the rank k available using the SVD of **X.** The k singular vectors are those sorted from the largest singular value to the smallest with each singular value (and vector) capturing sequentially decreasing variance (information) from **X**. Said another way, PCR estimates **b** by minimizing the least-squares criterion $\|\mathbf{X}_d\mathbf{b}-\mathbf{y}\|_2$ where $\mathbf{X}_d$ denotes **X** projected onto d basis vectors and $\|\bullet\|_2$represents the L_2 vector 2-norm.

A relationship often used to express computation of the PCR model vector based on d basis vectors is

$$\begin{aligned}\hat{\mathbf{b}} &= \mathbf{X}^{+}\mathbf{y}\\ &= \left(\mathbf{V}_d\Sigma_d^{-1}\mathbf{U}_d^T\right)\mathbf{y}\\ &= \sum_{i=1}^{d}\frac{\mathbf{u}_i^T\mathbf{y}}{\sigma_i}\mathbf{v}_i\\ &= \sum_{i=1}^{k}\beta_i\mathbf{v}_i\\ &= \mathbf{V}\beta\end{aligned} \tag{2}$$

where the subscript d is left off of $\hat{\mathbf{b}}$ and β_i denotes a scalar value for the weight given to a particular singular vector, i.e. the top d vectors have non-zero weights and the k - d weights are zero. More generally, all of the LPCR, CPCR, PCR, PLS, and RR regression vectors can be expressed by

$$\begin{aligned}\hat{\mathbf{b}} &= \sum_{i=1}^{k}\beta_i\mathbf{v}_i\\ &= \mathbf{V}\beta.\end{aligned} \tag{3}$$

As with PCR, the CPCR or LPCR methods have non-zero weights for those $\mathbf{v}_i$ vectors being used and zero for the others, i.e. the singular vectors with zero weights span the null spaces of $\mathbf{X}^T$. The PLS and PCR weight values for all k singular vectors depend on respective tuning parameter values. There is nothing particular about the **V** basis set and model regression vectors can be expressed as a linear combination of other basis set expansions [34 and references therein, 55].

An alternative to Equation (3) to characterize LPCR, CPCR, TPCR, PLS, and RR that ultimately reduces to Equation (3) is

$$\begin{aligned}\hat{\mathbf{b}} &= \mathbf{VF}\Sigma^{-1}\mathbf{U}^T\mathbf{y}\\ &= \sum_{i}^{k} f_i\frac{\mathbf{u}_i^T\mathbf{y}}{\sigma_i}\mathbf{v}_i\\ &= \mathbf{V}\beta\end{aligned} \tag{4}$$

where **F** signifies a $k\times k$ diagonal matrix of f_i filter values [47,53]. For LPCR, CPCR, and PCR, $f_i = 1$ for the $\mathbf{v}_i$ singular vectors retained and zero for the rest. For RR and PLS, the respective filter value ranges for all k vectors are $0\leq f_i\leq 1$ and $0\leq f_i\leq\infty$. Equation (4) expresses an alternative to characterize differences between calibration methods by inspecting filter values. Only the weight values for $\boldsymbol{\beta}$ are evaluated in this paper.

2.3. TR with the L_1 norm for singular vector selection; LPCR

The LPCR method is a variant of the TR method commonly known as LASSO. For RR, the L_2 norm penalty is used on the regression vector in the minimization expression $\min\left(\|\mathbf{Xb}-\mathbf{y}\|_2^2+\lambda^2\|\mathbf{b}\|_2^2\right)$ where λ denotes the RR tuning parameter. With LASSO, the L_1 vector norm (1-norm) penalty is used on the regression vector forming the minimization expression $\min\left(\|\mathbf{Xb}-\mathbf{y}\|_2^2+\tau\|\mathbf{b}\|_1\right)$ where τ represents the LASSO tuning parameter and $\|\cdot\|_1$signifies the L_1 norm. By using the L_1 norm, variables (wavelengths with spectral data) are selected to form sparse models. Instead of selecting wavelengths, TR is modified to select singular vectors from the SVD of **X.** The minimization expression for LPCR is

$$\min\left(\|\mathbf{U}\alpha-\mathbf{y}\|_2^2+\tau\|\alpha\|_1\right) \tag{5}$$

where $\boldsymbol{\alpha}$ symbolizes the $k\times 1$ model vector with non-zero coefficients for those basis vectors being selected and essentially zero for the remaining coefficients as with a wavelength selected LASSO model vector **b**. The terms in expression (5) stem from using the SVD on **X** and rewriting Equation (1) as $\mathbf{y}=\mathbf{U\Sigma V}^T\mathbf{b}$. Terms are combined to form $\mathbf{y}=\mathbf{U}(\mathbf{\Sigma V}^T\mathbf{b})=\mathbf{U}\boldsymbol{\alpha}$. An estimated singular vector basis set selected model vector solution $\hat{\boldsymbol{\alpha}}$ from expression (5) can be converted to a full wavelength model by $\hat{\mathbf{b}}=\mathbf{V\Sigma}^{-1}\hat{\boldsymbol{\alpha}}$ and then used for prediction in the usual way.

In this paper, the process for L_1 norm basis set selection first obtains the model number (τ) at the minimum of the mean RMSECV across LMOCV using a LASSO algorithm with Expression (5). At this model number, the absolute model vector coefficient values in the mean model vector $\hat{\boldsymbol{\alpha}}$ for the selected basis vectors from **U** across the LMOCVs are sorted from largest to smallest. The method of PCR is then used step-wise with these L_1 selected basis vectors sequentially added to form the LPCR model vectors and L-curve based graphical diagnostics.

2.4. PCR singular vector selection based on correlation; CPCR

For CPCR, the SVD singular vectors are ordered by decreasing absolute correlations between each $\mathbf{u}_i$ vector and **y**. The absolute correlation avoids negative correlations from mean centering (if used). Statistical testing or a threshold value on the correlation can be used [7–9,12–14]. In this paper, the singular vectors are ordered in sequence with decreasing absolute correlation. The basis vectors are then sequentially included to form the corresponding CPCR L-curve based plots.

2.5. Model updating

Calibration model updating involves updating a multivariate calibration model formed in primary conditions to predict samples measured in the secondary conditions. The approach used in this paper is a variant of TR where the primary samples are augmented with a few samples from the secondary conditions [22]. A weighting parameter (η) is used to emphasize a small set of samples from the secondary conditions. The minimization is expressed as $\min\left(\|\mathbf{Xb}-\mathbf{y}\|_2^2+\lambda^2\|\mathbf{b}\|_2^2+\eta^2\|\mathbf{Mb}-\mathbf{y_M}\|_2^2\right)$ for the model

$$\begin{pmatrix}\mathbf{y}\\ 0\\ \eta\mathbf{y_M}\end{pmatrix}=\begin{pmatrix}\mathbf{X}\\ \lambda\mathbf{I}\\ \eta\mathbf{M}\end{pmatrix}\mathbf{b} \tag{6}$$

with solution $\hat{\mathbf{b}}=\left(\mathbf{X}^T\mathbf{X}+\lambda^2\mathbf{I}+\eta^2\mathbf{M}^T\mathbf{M}\right)^{-1}\left(\mathbf{X}^T\mathbf{y}+\eta^2\mathbf{M}^T\mathbf{y_M}\right)$ where the **M** and $\mathbf{y_M}$ are the updating set of spectra and reference values for the secondary conditions and **I** is the identity matrix. With this model updating approach, model quality measures for the secondary conditions need to be

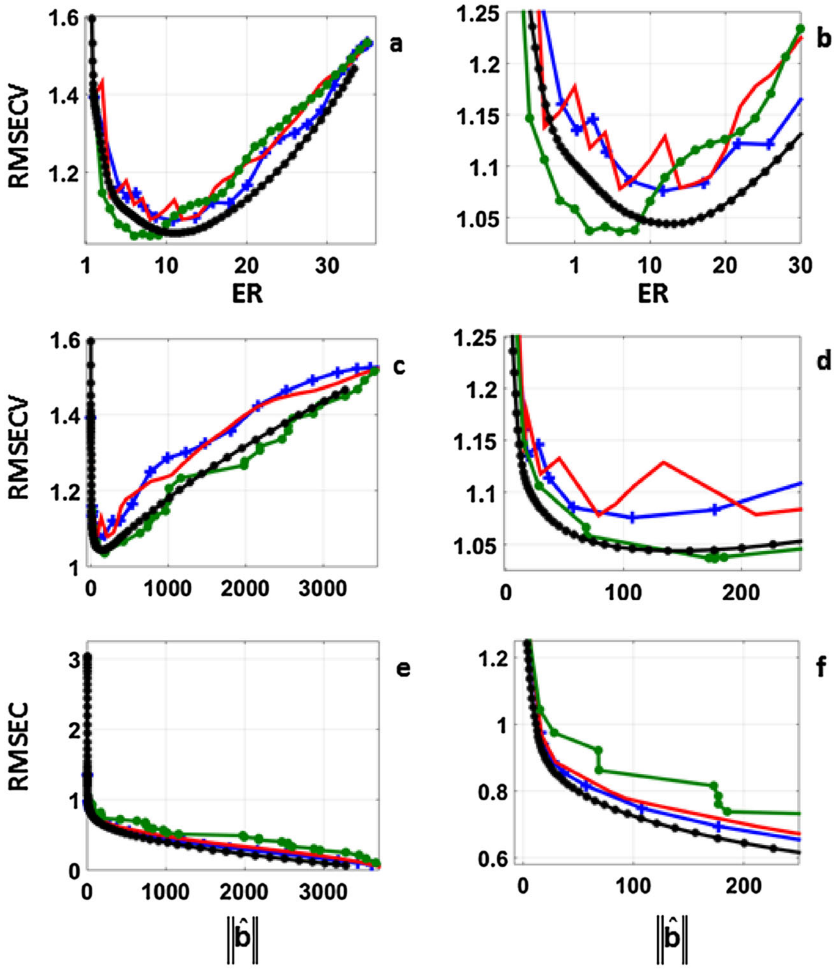

Figure 1. Soy calibration mean RMSECV curves plotted against (a) ER and (c) model L_2 norm. The corresponding expansions of (a) and (c) are in (b) and (d). The mean RMSEC is plotted against the model L_2 norm in (e) with an expansion in (f). LPCR (green circles), PCR (sold red line), PLS (blue plus sign), and RR (black asterisk).

assessed, such the RMSE of the samples forming **M** (RMSEM) and others.

To integrate LPCR into model updating, Equation (6) is modified to

$$\begin{pmatrix} \mathbf{y} \\ \eta \mathbf{y}_{\mathbf{M}} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \eta \mathbf{M} \end{pmatrix} \mathbf{b} \tag{7}$$

that can be expressed as $\mathbf{y}' = \mathbf{X}'\mathbf{b}$ where the two prime symbols indicate the augmented arrays on Equation (7). This equation is then solved using the LPCR process with the L_1 norm as noted in Expression (5). Equation (7) is also solvable by CPCR, PCR, and PLS.

When using LPCR with Equation (7), there are two tuning parameters, which singular vectors (τ in Expression (5)) and a value for η. While L-curve approaches have been evaluated for selecting unique values for the two tuning parameters [22], a recent consensus approach [56] is used to select models over a range of tuning parameter values. Thus, a family of models instead of "a model" for calibration updating are selected. In consensus or ensemble modeling,

Table I. Soy results at selected models based on minimum RMSECV and L-curves[a]

Method	Tuning parameter[b]	ER	$\|\hat{\mathbf{b}}\|_2$	RMSEC	RMSECV	R^2	Slope	Intercept
LPCR	1, 2, 3, 6, 8, 12(1,3,8)	6(3)	172(29.5)	0.82(0.89)	1.04(1.12)	0.91(0.88)	0.92(0.89)	1.01(1.22)
PCR	8(6)	8(6)	79(27.9)	0.80(0.98)	1.08(1.11)	0.89(0.89)	0.90(0.89)	1.01(1.32)
PLS	7(5)	10(7)	107(36.5)	0.75(0.86)	1.08(1.12)	0.89(0.89)	0.92(0.90)	0.95(1.16)
RR	$\lambda_{68} = 0.015(\lambda_{56} = 0.070)$	10(6)	107(26.3)	0.72(0.87)	1.05(1.09)	0.96(0.89)	0.92(0.89)	1.03(1.29)

[a] Values in parentheses are from L-curve selected models.
[b] SVD basis vectors for LPCR, number of SVD basis vectors for PCR, number of PLS LVs, and ridge value.

a sample is predicted with a collection of models and a composite (fused) prediction is formed [2,57–61]. The typical process forms multiple models by random sampling across samples (bagging), variables (random subspace method), or both. The consensus approach from previous work [56] is used in this study. Briefly, two-dimensional landscapes of the same model quality measures used in simple L-curve type plots are used as the tuning parameters vary. These landscapes are assessed visually to determine tradeoff regions. Threshold values are then set for the model quality

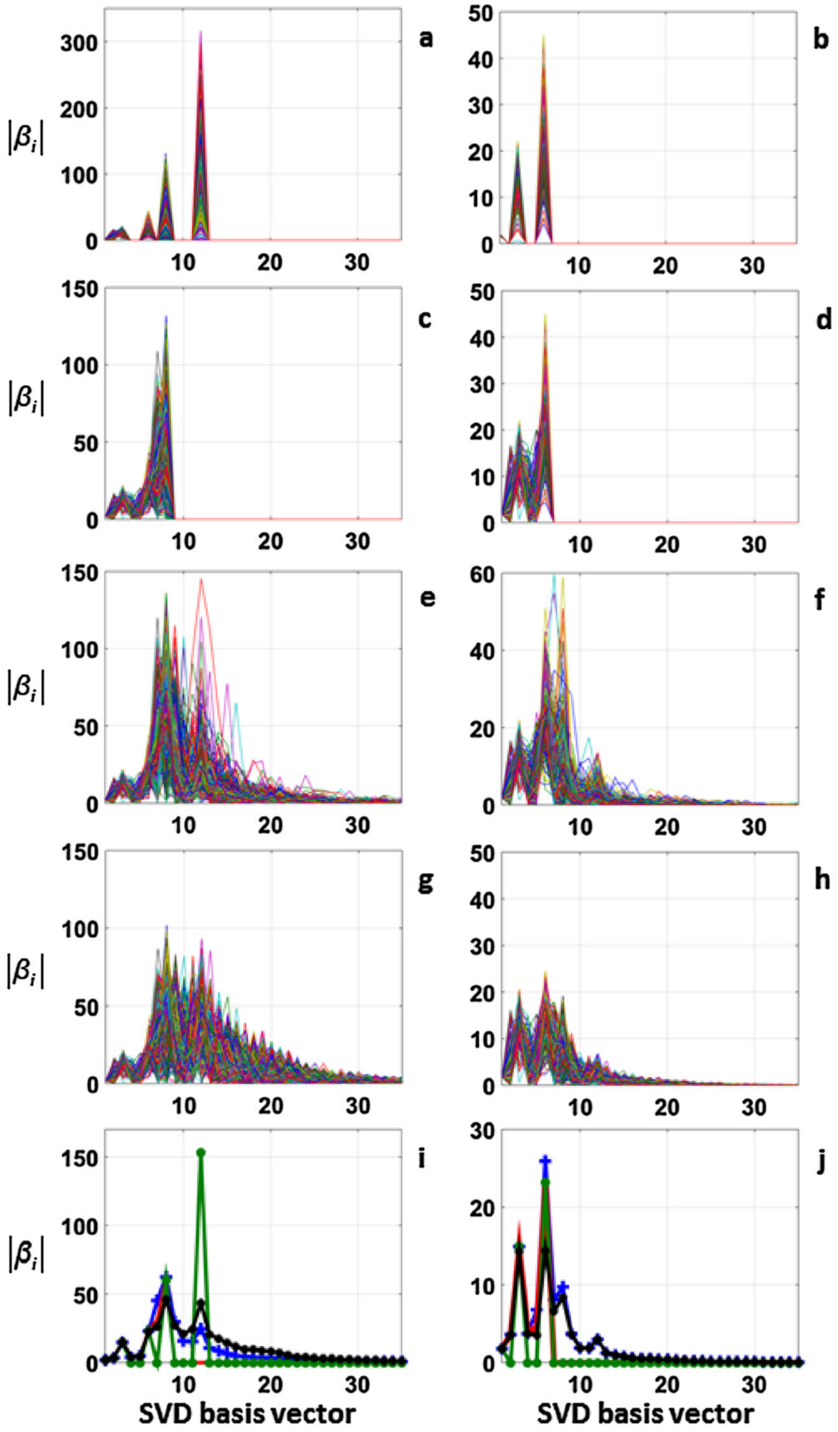

Figure 2. All respective CV soy model β vectors for the SVD basis set (Equation (4)) at the models forming the mean minima RMSECV values for (a) LPCR, (c) PCR, (e) PLS, and (g) RR. Plotted in (i) are the mean weight values across all the CVs for LPCR (green circles), PCR (sold red line), PLS (blue plus sign), and RR (black asterisk). The corresponding plots in (b), (d), (f), (h), and (j) are at the models in the corner region of the L-curve in Figures 1e and f.

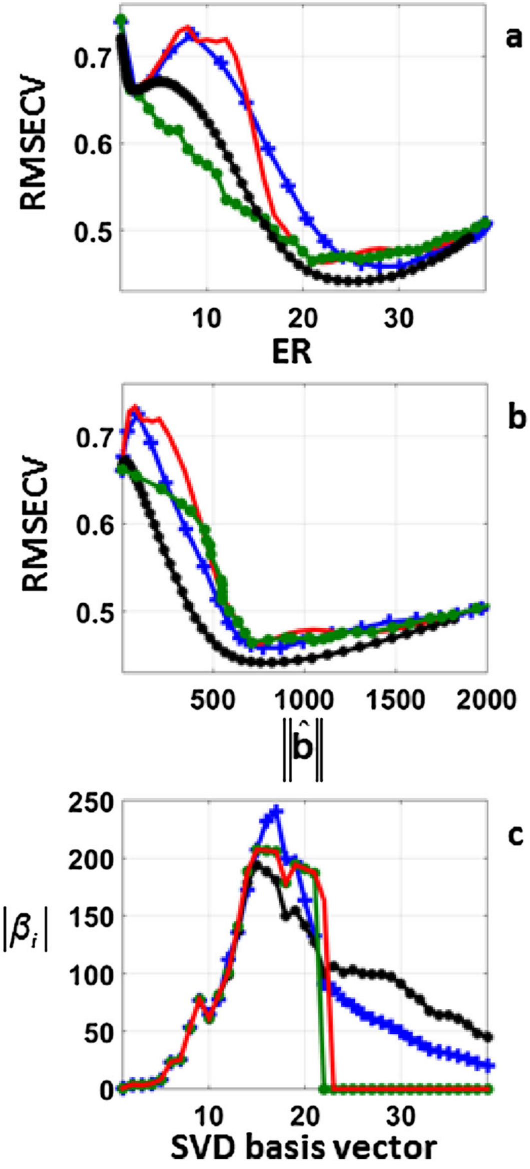

Figure 3. Wheat calibration mean RMSECV curves plotted against (a) ER and (b) model L_2 norm. Plotted in (c) are the mean weight values across all the CVs. LPCR (green circles), PCR (sold red line), PLS (blue plus sign), and RR (black asterisk).

measures relative to values in optimal tradeoff regions. Models passing thresholds are then used for prediction of new samples from the secondary conditions.

3. EXPERIMENTAL

3.1. Software

The MATLAB programs used in this study were written by the authors using MATLAB 2010b (The MathWorks, Natick, MA).

3.2. Models and CVs

For each data set, LMOCV was used with 500 data splits using 60% of the samples as the calibration set and the remaining samples as the validation set. On each LMOCV split, ER is calculated for each modeling method across all respective tuning parameters using normally distributed random noise added to $\mathbf{y}$ 300 times (N) with $\boldsymbol{\delta}$ set to 0.5. Eigenvector weights are also calculated on each LMOCV split for each respective tuning parameter. Mean model assessment values of the 500 LMOCVs are used to select tuning parameters as well as the values reported for plots and tabulated results. For RR and model updating, the λ and η values for all methods are the same and ranged from 112 to 10^{-3} with 100 values. On each CV data split, the calibration set is mean centered and the validation is centered to the calibration mean. For the calibration updating study with the pharmaceutical tablet data sets, additional specifics are given in the data set description.

3.3. Soy data

The soy data set consists of near infrared (NIR) spectra for 60 samples measured from 1100 to 2500 nm at 4-nm intervals for 350 wavelengths [62]. Protein content is the analyte.

3.4. Wheat data

There are 87 wheat samples measured in the NIR from 1100 to 2500 nm at 10-nm intervals for 140 wavelengths [63]. Protein content is the analyte.

3.5. Corn data

The 80 samples are measured in the NIR from 1100 to 2500 nm at 2-nm intervals for 700 wavelengths [64]. Moisture content is the analyte and spectra measured on m5 are used.

3.6. Tablet data

The pharmaceutical tablet data set consists of 310 Escitolopram tablets measured in the range of 7400–10 507 cm^{-1} for a total of 404 values [65]. Tablets are subdivided into four types (type

Table II. Wheat results at selected models based on minimum RMSECV

Method	Tuning parameter[a]	ER	$\|\hat{\mathbf{b}}\|_2$	RMSEC	RMSECV	R^2	Slope	Intercept
LPCR	1–21	21	710	0.19	0.46	0.73	0.86	1.61
PCR	22	22	739	0.18	0.46	0.74	0.87	1.61
PLS	14	28	773	0.16	0.46	0.74	0.88	1.43
RR	$\lambda_{86}=0.001$	25	760	0.14	0.44	0.75	0.85	1.74

[a] SVD basis vectors for LPCR, number of SVD basis vectors for PCR, number of PLS LVs, and ridge value.

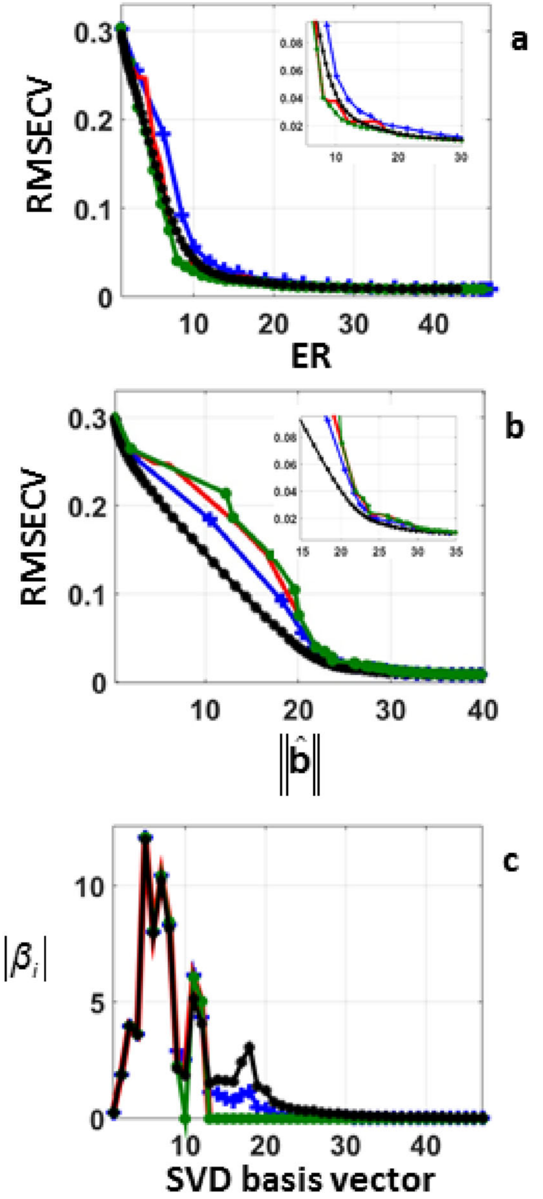

Figure 4. Corn calibration mean RMSECV curves plotted against (a) ER and (b) model L_2 norm. Enlargements are shown as insets. Plotted in (c) are the mean weight values across all the CVs. LPCR (green circles), PCR (sold red line), PLS (blue plus sign), and RR (black asterisk).

1, type 2, type 3, and type 4) based on total tablet weights of 90, 125, 188, and 250 mg. Because tablet types have different total weights, respective tablet types have different shapes and sizes with tablet thicknesses ranging from 2.9 to 4.3 mm. There are 30 tablets for each batch tablet type combination. As used in previous calibration updating work, tablets from laboratory and full batches are used for the calibration updating study [66]. The 60 lab batch tablets of types 1 and 2 are used as a fixed primary calibration set (**X** and **y** in Equations (6) and (7)). Full batch tablet types 1 and 2 are set to the secondary condition with eight tablets (four of each type) used for the updating set and the remaining 52 tablets for validation. One hundred random CVs were used to form the updating ($\mathbf{M}$ and $\mathbf{y}_\mathbf{M}$ in Equations (6) and (7)) and validation sets for the secondary full batch tablets. Because there is only one primary calibration set, it is mean centered to its mean only once. On each CV split of the secondary tablets, the updating set is mean centered and the secondary validation samples are centered to this mean.

4. RESULTS AND DISCUSSION

Results from using CPCR for forming a subset of selected basis vectors are not shown for the calibration modeling portion of this study. The calibration behavior is the same as LPCR up to the minimum RMSECV from the LMOCV (which is used to select the subset of singular vectors). After the minimum RMSECV, the vectors selected by LPCR and CPCR deviated; probably because of the random chance correlations from the noise in the remaining basis vectors. Conversely, for the calibration updating, LPCR and CPCR behaved differently, and results are presented for both processes. As noted previously, basis vectors were not tested for significance and similarly, statistical testing [67] was not performed on the model quality measures for the different calibration processes. Instead, the model quality measures are tracked in conjunction with basis vectors weights.

4.1. Soy data set for calibration

Mean RMSCV and model complexity measures are plotted in Figures 1a and c. Expansion of the plots are shown in Figures 1b and d. Using $\|\hat{\mathbf{b}}\|_2$ or the model ER allows comparison between the methods on one plot compared to separate plots with the tuning parameter on the x-axis. At the RMSECV minima, the ERs indicate the LASSO L_1 norm sorted LPCR model has the smallest ER indicating the least complexity. However, with the $\|\hat{\mathbf{b}}\|_2$ as the complexity measure, the LPCR and RR models appear more similar. The PCR models have erratic RMSECV behavior. As

Table III. Corn results at selected models based on minimum RMSECV

Method	Tuning parameter[a]	ER	$\|\hat{\mathbf{b}}\|_2$	RMSEC	RMSECV	R^2	Slope	Intercept
LPCR	1–9,11,12	11	23.7	0.02	0.02	1.00	0.99	0.06
PCR	12	12	23.9	0.02	0.02	1.00	1.00	0.04
PLS	9	17	23.1	0.02	0.02	1.00	1.00	0.02
RR	$\lambda_{68} = 0.008$	10	23.9	0.01	0.02	1.00	0.99	0.10

[a] SVD basis vectors for LPCR, number of SVD basis vectors for PCR, number of PLS LVs, and ridge value.

expected, RR has the smoothest plots regardless of the model complexity measure.

Selecting the models at the minimum RMSECV values, some model quality measures are tabulated in Table I. While the RMSECV for LPCR is the smallest, the model vector L_2 norm is the largest. The values listed in Table I reveal that tradeoffs exists between the different modeling methods when using the minimum RMSECV to select the final model.

Table IV. Tablet threshold model quality measure values for model updating

Measure	Minimum	Maximum
R^2 **X**,**y**	0.75	0.90
Slope **X**,**y**	0.95	0.99
Intercept **X**,**y**	0.05	0.50
R^2 **M**,$\mathbf{y_M}$	0.98	0.99
Intercept **M**,$\mathbf{y_M}$	0.05	0.50
Slope **M**,$\mathbf{y_M}$	0.95	0.99

Shown in the first column of Figure 2 are all the CV data splits for the respective model **β** vectors from Equation (4) at the models forming the minima RMSECV values using the corresponding SVD basis vectors. Plotted in Figure 2i are the mean weight values across all the CV. From the plots in the first column of Figure 2, it is clear that LPCR has selected a subset of basis vectors and as expected, the PLS and RR models can make use of all the rank *k* SVD basis vectors. While only the six SVD basis vectors 1, 2, 3, 6, 8, and 12 are used for LPCR, the method utilizes the 12th basis vector compared to PCR stopping at the 8th basis vector. While TPCR stops at the 8th basis vector, PLS and RR use more basis vectors with RR applying a greater weight than PLS at the later basis vectors. From the weight values plotted for all the data splits, all the methods show erratic weighting behavior except LPCR. The large weight given to the 12th SVD basis vector for the LPCR model explains the large L_2 norm.

Instead of using the minima in RMSECV plots, the bias/variance tradeoff is assessed at points before the respective RMSECV minima using the mean RMSEC, RMSECV,

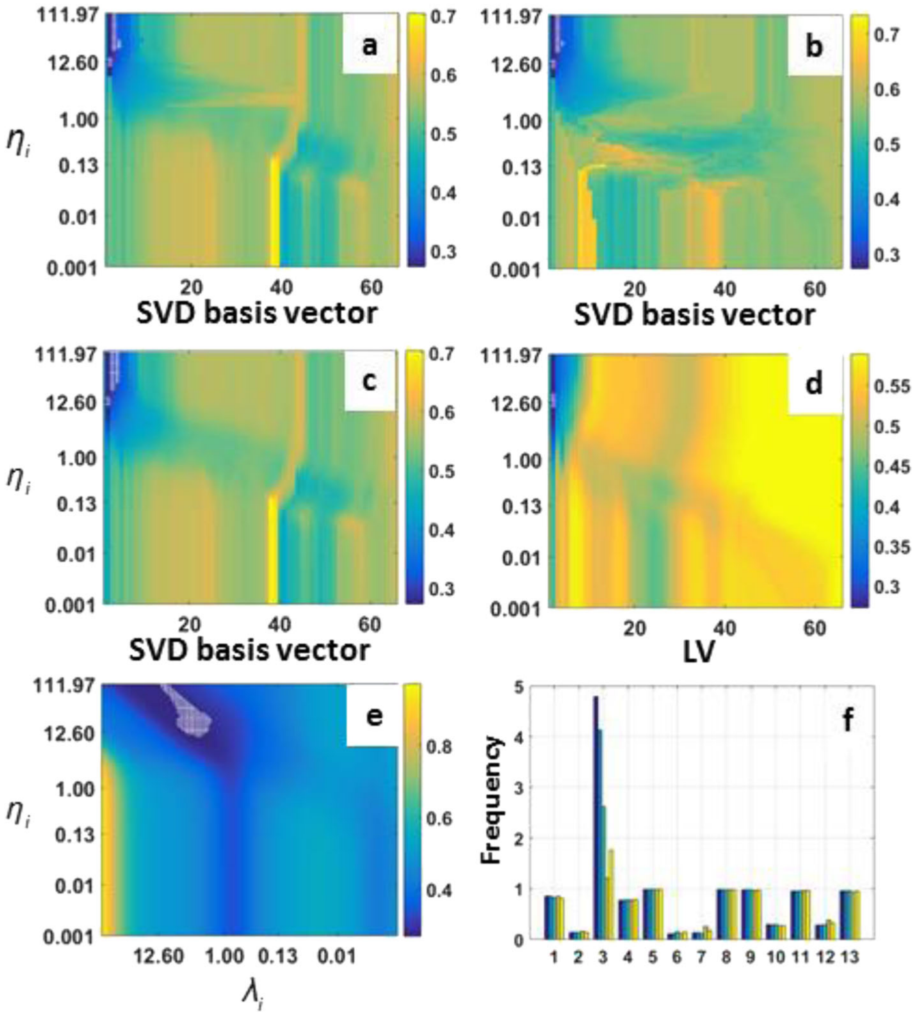

Figure 5. Tablet RMSECV landscapes for (a) LPCR, (b) CPCR, and (c) PCR, (d) PLS, and (e) TR. White plus signs correspond to models passing thresholds provided in Table IV. In (f) is a bar plot of the mean model quality measures in Tables V and VI with 1 = RMSEC, 2 = RMSEM, 3 = $\left\|\hat{\mathbf{b}}\right\|_2$, 4 = R^2 (**X**,**y**), 5 = R^2 (**M**,$\mathbf{y_M}$), 6 = intercept (**X**,**y**), 7 = intercept (**M**,$\mathbf{y_M}$), 8 = slope (**X**,**y**), 9 = slope (**M**,$\mathbf{y_M}$), 10 = RMSECV, 11 = R^2 CV, 12 = intercept CV, 13 = slope CV. The order of the bars are LPCR, CPCR, PCR, PLS, and TR.

Table V. Tablet LPCR, CPCR, and PCR results at selected updated models based on thresholds in Table IV

Model measure	LPCR (21 models)			CPCR (19 models)			PCR (30 models)		
	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
RMSEC	0.79	0.86	0.88	0.78	0.86	0.88	0.78	0.83	0.87
RMSEM	0.125	0.14	0.17	0.12	0.14	0.17	0.13	0.14	0.17
$\left\|\hat{\mathbf{b}}\right\|_2$	1.15	4.79	8.64	1.15	4.13	8.64	1.15	2.62	3.57
R^2 **X**,**y**	0.77	0.77	0.80	0.77	0.77	0.80	0.77	0.79	0.80
R^2 **M**,$\mathbf{y_M}$	0.98	0.99	0.99	0.98	0.98	0.99	0.98	0.99	0.99
Intercept **X**,**y**	0.06	0.11	0.24	0.06	0.11	0.30	0.08	0.16	0.30
Intercept **M**,$\mathbf{y_M}$	0.09	0.13	0.33	0.09	0.14	0.33	0.08	0.12	0.33
Slope **X**,**y**	0.96	0.98	0.99	0.96	0.98	0.99	0.96	0.98	0.99
Slope **M**,$\mathbf{y_M}$	0.95	0.98	0.99	0.95	0.98	0.99	0.95	0.98	0.99
RMSECV	0.27	0.30	0.32	0.27	0.29	0.32	0.27	0.30	0.31
R^2	0.95	0.96	0.96	0.95	0.96	0.96	0.95	0.96	0.96
Intercept	0.25	0.28	0.46	0.25	0.29	0.46	0.26	0.29	0.46
Slope	0.93	0.96	0.96	0.93	0.96	0.96	0.93	0.96	0.96
η	11.22	46.18	111.9	11.22	48.23	111.9	11.22	55.20	111.9
No. SVD basis vectors	2	3	4	2	2.9	4	2	3.3	4

and the $\left\|\hat{\mathbf{b}}\right\|_2$ plotted in Figures 1c–f. The corner regions are at smaller $\left\|\hat{\mathbf{b}}\right\|_2$ and ER values than at the RMSECV minima indicating the minimum RMSECV models are probably overfitted. From Figure 1, it is observed that the L-curves with RMSEC closely mimic the RMSECV plots up to the RMSECV minima as is the usual case. Listed in Table I are the model quality measures for models in the corner region. Again, there are tradeoffs between the different calibration methods but all are essentially the same except PCR which has a small ER value.

Plotted in the second column of Figure 2 are the corresponding SVD basis vectors weights for all the CV data splits and the mean weight values across all the CV. From these plots in Figure 2, it is observed that LPCR selects basis vectors 1, 3, and 8. The PCR models again tend to be more erratic in the weighting behavior. The PLS and RR weight later basis vectors, but not as much as with the models at the minima RMSECV.

4.2. Wheat data set for calibration

Shown in Figure 3 are the mean RMSCV and model complexity measures. From these plots, the models selected are at the minima RMSECV values because of the agreement of the minima corresponding the corner to the RMSEC L-curves. The mean SVD basis weight values (β_i) at the selected models across all the CV splits are shown in Figure 3. The only unique trend observed with this data set is that LPCR selected a top-down ordered subset of basis vectors stopping at basis vector 21 while TPCR stops at the 22nd basis vector. Values tabulated in Table II show that all the models are essentially predicting the same.

4.3. Corn data set for calibration

The plots of mean RMSCV values and model complexity measures displayed in Figure 4 reveal no minima and the curves are shaped as RMSEC L-curves, i.e. the corner regions of the RMSEC L-curve complements the corner region of the RMSECV plot. This is not uncommon and hence, using the corner region of the bias/variance tradeoff for the calibration set is useful in this situation. The mean SVD basis weights presented in Figure 4 show that all the calibration methods use the same basis vectors (but weighted differently as with the other data sets) with the following exceptions: LPCR uses basis vectors 1–12 except the 10th, PCR uses 1–12, and PLS and RR use most of the rank k basis vectors with emphasis on the 1–12 basis vectors and slight weight enhancements on basis vectors 17 and 18 for PLS. The small variations in the model measures of quality tabulated in Table III indicate that the models have no distinctive differences between the methods.

4.4. Tablet data set for model updating

As noted in section 2.5 on Model updating, the approach used in this paper to select the two tuning parameter values for each

Table VI. Tablet PLS and TR results at selected updated models based on thresholds in Table IV

Model measure	PLS (5 models)			TR (121 models)		
	Min.	Mean	Max.	Min.	Mean	Max.
RMSEC	0.84	0.85	0.86	0.74	0.81	0.84
RMSEM	0.16	0.16	0.16	0.12	0.14	0.17
	1.21	1.21	1.21	1.01	1.76	2.84
R^2 **X**,**y**	0.77	0.77	0.77	0.78	0.79	0.81
R^2 **M**,$\mathbf{y_M}$	0.98	0.98	0.98	0.98	0.99	0.99
Intercept **X**,**y**	0.05	0.12	0.19	0.05	0.15	0.34
Intercept **M**,$\mathbf{y_M}$	0.21	0.25	0.31	0.10	0.17	0.33
Slope **X**,**y**	0.97	0.98	0.99	0.95	0.98	0.99
Slope **M**,$\mathbf{y_M}$	0.95	0.96	0.97	0.95	0.97	0.98
RMSECV	0.27	0.27	0.27	0.27	0.28	0.29
R^2	0.96	0.96	0.96	0.96	0.96	0.96
Intercept	0.33	0.38	0.44	0.27	0.32	0.46
Slope	0.94	0.94	0.95	0.93	0.95	0.96
η	11.22	14.31	17.78	11.22	33.54	112.0
No. LVs (PLS) λ (TR)	2	2	2	1.78	4.34	11.22

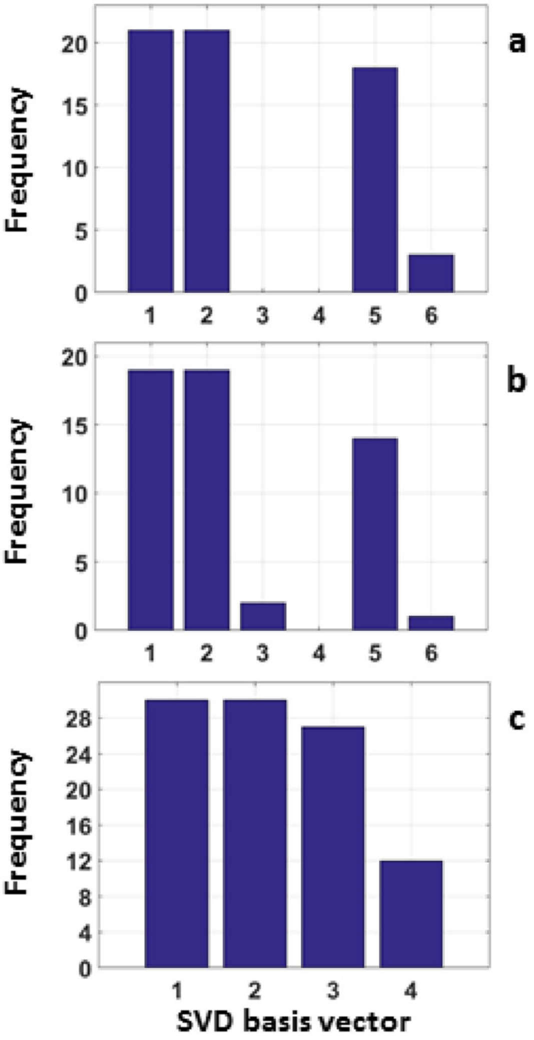

Figure 6. Histograms of the selected SVD basis vectors for (a) LPCR, (b) CPCR, and (c) PCR tablet models listed in Table V that passed the thresholds listed in Table IV.

modeling method is to set up thresholds on measures of model quality [56]. This process results in a collection of acceptable models. While threshold values are data set dependent, natural target values are respectively 1, 1, and 0 for R^2, slope, and intercept from plotting predicted values against reference values. In order to set up thresholds, landscapes of these model quality measures (and others) for the primary samples and secondary samples are formed, i.e. images of the measures as the tuning parameters vary (not shown). These landscapes are inspected for regions with acceptable bias/variance tradeoffs. Listed in Table IV are the threshold values used for the Tablet data set. The final prediction for a sample is the mean prediction from the collection of models, but other combinations of the prediction values can be used.

Plotted in Figure 5 are RMSECV landscapes using LPCR, CPCR, PCR, PLS, and TR. The landscapes are similar but because of the discreetness in the PLS and PCR tuning parameters (LVs and SVD basis vectors), the optimal model region is sharp. For TR with a continuous tuning parameter (λ), landscape is shaped like a bowl in the optimal model region.

Shown in Table V are the mean model quality measures from the collection of LPCR, CPCR, and PCR passing the thresholds. Also listed are the number of models passing the thresholds and the ranges for the model quality measures. The PLS and TR values for the same measures of models passing thresholds are listed in Table VI. All respective models passing the thresholds are shown on the RMSECV landscapes in Figure 5. From the values tabulated in the Tables V and VI, it appears that except for the L_2 norms of the model vectors, one method does not outperform another and all five approaches are equivalent. This equivalency (except model vector L_2 norms) is characterized by the bar plot in Figure 5f of the mean values listed in Tables V and VI. The model L_2 norms are largest for LPCR and CPCR due the some of the selected models having large values. The other methods generally end up with models with a more constant model L_2 norm.

A goal of LPCR and CPR is to allow selection of a subset of SVD basis vectors not necessarily in the top-down order. Shown in Figure 6 are the histograms of the basis vectors selected for the LPCR, CPCR, and PCR models passing the thresholds. Most of the LPCR and CPCR models use SVD basis vectors 1, 2, and 5. The differences between the two methods is that LPCR does not use basis vector 3 while CPCR does for a few of the models. Basis vectors for PCR are mostly 1 through 3 with some models including basis vector 4. Even though there is a mix of basis vectors between LPCR, CPCR, and PCR, the values in Table V and the bar plot in Figure 5f indicate there are no noteworthy differences between the methods except for the L_2 norms of the model vectors.

5. CONCLUSION

The LPCR and CPCR calibration and model updating methods have the flexibility to select a subset of SVD basis vectors from total possible. However, even if the methods actually form models with a non top-down ordered subset, the model quality measures evaluated in this paper do not show any notable advantages. The model vector L_2 norms were found to usually be larger than the PCR, PLS, RR, and TR models when a non top-down set is selected by LPCR and CPCR. Thus, with LPCR and CPCR, there is the potential for larger predication variances than the other three methods because of the larger model vector L_2 norms. A possible use for using LPCR or CPCR is to first project a data set with the basis vectors found best for prediction with LPCR or CPCR and then apply PLS or RR. This preprocessing step by LPCR or CPCR may remove basis vectors not associated with analyte.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. CHE-1111053 (co-funded by the MPS Chemistry and the OCI Venture) and is gratefully acknowledged by the authors. Summer support from National Science Foundation Award 1156879 is acknowledged by Trevor O'Loughlin.

REFERENCES

1. Næs T, Isaksson T, Fern T, Davies T. *A User Friendly Guide to Multivariate Calibration and Classification*, NIR Publications: Chichester, UK, 2002.
2. Hastie TJ, Tibshirani RJ, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, (2nd edn)Springer-Verlag: New York, 2009.

3. Kalivas JH. Calibration methodologies.In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis Vol. 3*, (eds) eds-in-chief. Elsevier: Amsterdam, 2009; 1–32.
4. Kalivas JH, Palmer J. Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters. *J. Chemom.* 2014; **28**: 347–357.
5. Lorber A. Error propagation and figures of merit for quantification by solving matrix equations. *Anal. Chem.* 1986; **58**: 1167–1172.
6. Olivieri AC, Faber NM, Ferré J, Boqué R, Kalivas JH, Mark H. Quantifying selectivity in spectrophotometric multicomponent analysis. *Pure & Appl. Chem.* 2006; **78**: 633–661.
7. Jolliffe IT. *Principal Component Analysis*, Springer-Verlag: New York, 1986.
8. Myers RH. *Classical and Modern Regression with Applications*, (2nd edn)Duxbury: Pacific Grove, 1990.
9. Downey G, Robert P, Bertrand D, DeVaux MF. Dried grass silage analysis by NIR reflectance spectroscopy-a comparison of stepwise multiple linear and principal component techniques for calibration development on raw and transformed spectral data. *J. Chemom.* 1989; **3**: 397–407.
10. Sutter JM, Kalivas JH, Lang PM. Which principal components to utilize for principal component regression. *J. Chemom.* 1992; **6**: 217–225.
11. Barros AS, Rutledge DN. Genetic algorithm applied to the selection of principal components. *Chemom. Intell. Lab. Syst.* 1998; **40**: 65–81.
12. Verdú-Andrés J, Massart DL. Comparison of prediction- and correlation-based methods to select the best subset of principal components for principal component regression and detect outlying objects. *Appl. Spectrosc.* 1998; **52**: 1425–1434.
13. Sun J. A correlation principal component regression analysis of NIR data. *J. Chemom.* 1995; **9**: 21–29.
14. Fairchild SZ, Kalivas JH. PCR eigenvector selection based on correlation relative standard deviations. *J. Chemom.* 2001; **15**: 615–625.
15. Davis AMC. A better way to of doing principal component regression. *Spectrosc. Eur.* 1995; **7**: 36–38.
16. Tikhonov AN. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* 1963; **4**: 1035–1038.
17. Claerbout JF, Muir F. Robust modeling with erratic data. *Geophysics* 1973; **38**: 826–844.
18. Taylor HL, Banks SC, McCoy JF. Deconvolution with the l_1 norm. *Geophysics* 1979; **44**: 39–52.
19. Levy S, Fullager PK. Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics* 1981; **46**: 1235–1243.
20. Santosa F, Symes W. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* 1986; **7**: 1307–1330.
21. Tibshirani R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B* 1996; **58**: 267–288.
22. Kalivas JH. Overview of two-norm (L_2) and one-norm (L_1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. *J. Chemom.* 2012; **26**: 218–230.
23. Jolliffe IT, Trendafilow NT, Uddin M. A modified principal component technique based on LASSO. *J. Comput. Graph. Stat.* 2003; **12**: 531–547.
24. Zou H, Hastie R, Tibshirani R. Sparse principal component analysis. *J. Comput. Graph. Stat.* 2006; **15**: 265–286.
25. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostat.* 2009; **10**: 515–534.
26. Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B Stat. Methodol.* 2010; **72**: 3–25.
27. Filzmoser P, Gschwandtner M, Todorov V. Review of sparse methods in regression and classification with application to chemometrics. *J. Chemom.* 2011; **26**: 42–51.
28. Hocking RR, Speed FM, Lynn MJ. A class of biased estimators in linear regression. *Technometrics* 1976; **18**: 425–437.
29. O'Sullivan F. A statistical perspective on ill-posed inverse problems. *Stat. Sci.* 1986; **1**: 502–527.
30. Marquardt DW. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* 1970; **12**: 591–612.
31. Gunst RF, Mason J. Biased estimation in regression: an evaluation using mean squared error. *J. Am. Stat. Assoc.* 1977; **72**: 616–628.
32. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993; **35**: 109–148.
33. Kalivas JH. Interrelationships of multivariate regression methods using eigenvector basis sets. *J. Chemom.* 1999; **13**: 111–132.
34. Kalivas JH. Basis sets for multivariate regression. *Anal. Chim. Acta* 2001; **428**: 31–40.
35. Kalivas JH, Green RL. Pareto optimal multivariate calibration for spectroscopic data. *Appl. Spectrosc.* 2001; **55**: 1645–1652.
36. Brown SD. Transfer of multivariate calibration models.In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis Vol. 3*, (eds)eds-in-chief. Elsevier: Amsterdam, 2009; 345–377.
37. Faber NM, Rajkó R. How to avoid over-fitting in multivariate calibration—the conventional validation approach and an alternative. *Anal. Chim. Acta* 2007; **595**: 98–106.
38. Wiklund S, Nilsson D, Eriksson L, Sjöström M, Wold S, Faber K. A randomization test for PLS component selection. *J. Chemom.* 2007; **21**: 427–439.
39. Wasim M, Brereton RG. Determination of the number of significant components in LC NMR spectroscopy. *Chemom. Intell. Lab. Syst.* 2004; **72**: 133–151.
40. Höskuldsson A. Dimension of linear models. *Chemom. Intell. Lab. Syst.* 1996; **32**: 37–55.
41. Denham MC. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *J. Chemom.* 2000; **14**: 351–361.
42. Li B, Morris J, Martin EB. Model selection for partial least squares regression. *Chemom. Intell. Lab. Syst.* 2002; **64**: 79–89.
43. Gómez-Carracedo MP, Andrade JM, Rutledge DN, Faber NM. Selecting the optimum number of partial least squares component for the calibration of attenuated total reflectance-mid-infrared spectra of undersigned kerosene samples. *Anal. Chim. Acta* 2007; **585**: 253–265.
44. Bauer F, Lukas MA. Comparing parameter choice methods for regularization of ill-posed problems. *Math. Comput. Simul.* 2011; **81**: 1795–1841.
45. Kalivas JH, Héberger K, Andries E. Using sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods. *Anal. Chim. Acta* 2015; **869**: 21–33.
46. Gowen AA, Downey G, Esquerre C, O'Donnell CP. Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *J. Chemom.* 2011; **25**: 375–381.
47. Hansen PC. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM: Philadelphia, PA, 1998.
48. Hansen PC. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev.* 1992; **34**: 561–580.
49. Stout F, Baines M, Kalivas JH. Impartial graphical comparison of multivariate calibration methods and the harmony/parsimony tradeoff. *J. Chemom.* 2006; **20**: 464–475.
50. van der Voet H. Pseudo-degrees of freedom for complex predictive models; an example of partial least squares. *J. Chemom.* 1999; **13**: 195–208.
51. Seipel HA, Kalivas JH. Effective rank for multivariate calibration methods. *J. Chemom.* 2004; **18**: 306–311.
52. Kalivas JH, Seipel HA. Erratum to Seipel HA, Kalivas JH. Effective rank for multivariate calibration methods. J. Chemom. 2004; 18: 306–311. *J. Chemom.* 2005; **19**: 64.
53. Andries E, Kalivas JH. Multivariate calibration leverages and spectral F-ratios via the filter factor representation. *J. Chemom.* 2010; **24**: 249–260.
54. Ye J. On measuring and correcting the effects of data mining and model selection. *J. Am. Stat. Assoc.* 1998; **93**: 120–131.
55. Green RL, Kalivas JH. Graphical diagnostics for regression model determinations with consideration of the bias/variance trade-off. *Chemom. Intell. Lab. Syst.* 2002; **60**: 173–188.
56. Shahbazikhah P, Kalivas JH. A consensus modeling approach to update a spectroscopic calibration. *Chemom. Intell. Lab. Syst.* 2013; **120**: 142–153.
57. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 1998; **20**: 832–844.
58. Ho TK. Decision combination in multiple classifier systems. *IEEE Trans. Pattern Anal. Mach. Intell.* 1994; **16**: 66–75.
59. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.* 1998; **20**: 226–239.
60. Tong W, Hong H, Fang H, Xie Q, Perkins R. Decision forests: combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* 2003; **43**: 525–531.

61. van Rhee AM. Use of recursion forests in the sequential screening process: consensus selection by multiple recursion trees. *J. Chem. Inf. Comput. Sci.* 2003; **43**: 941–948.
62. Bouveresse E, Hartmann C, Massart DL, Last IR, Prebble KA. Standardization of near-infrared spectrometric instruments. *Anal. Chem.* 1996; **68**: 982–990.
63. Kalivas JH. Two data sets of near infrared spectra. *Chemom. Intell. Lab. Syst.* 1997; **37**: 255–259.
64. Wise BM, Gallagher NB, PLS_Toolbox, Eigenvector Research, Manson, Washington. http://www.eigenvector.com/data/Corn/index.html, accessed July 2, 2015.
65. Dyrby M, Engelsen SB, Nørgaard L, Bruhn M, Lundsberg-Nielsen L. Chemometric quantitation of the active substance3 (containing C≡N) pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-IR Raman spectra. *Appl. Spectrosc.* 2001; **56**: 579–585. http://www.models.life.ku.dk/datasets accessed July 2, 2015.
66. Farrell J, Higgins K, Kalivas JH. Updating a near-infrared multivariate calibration model formed with lab-prepared pharmaceutical tablet types to new tablet types in full production. *J. Pharm. Biomed. Anal.* 2012; **61**: 114–121.
67. van der Voet H. Comparing the predictive accuracy of models using a simple randomization test. *Chemom. Intell. Lab. Syst.* 1994; **25**: 313–323.