Spectral Multivariate Calibration with Wavelength Selection Using Variants of Tikhonov Regularization

JOSHUA OTTAWAY, JOHN H. KALIVAS,* and ERIK ANDRIES

Department of Chemistry, Idaho State University, Pocatello, Idaho 83209 (J.O., J.H.K.); Department of Mathematics, Central New Mexico Community College, Albuquerque, New Mexico 87106 (E.A.); and Center for Advanced Research Computing, University of New Mexico, Albuquerque, New Mexico 87106 (E.A.)

Tikhonov regularization (TR) is a general method that can be used to form a multivariate calibration model and numerous variants of it exist, including ridge regression (RR). This paper reports on the unique flexibility of TR to form a model using full wavelengths (RR), individually selected wavelengths, or multiple bands of selected wavelengths. Of these three TR variants, the one based on selection of wavelength bands is found to produce lower prediction errors. As with most wavelength selection algorithms, the model vector magnitude indicates that this error reduction comes with a potential increase in prediction uncertainty. Results are presented for near-infrared, ultraviolet–visible, and synthetic spectral data sets. While the focus of this paper is wavelength selection, the TR methods are generic and applicable to other variable-selection situations.

Index Headings: **Tikhonov regularization; Wavelength selection; Variable selection; Multivariate calibration.**

INTRODUCTION

Multivariate calibration models used to analyze spectroscopic data have the general form

$$\mathbf{y} = \mathbf{Xb} \tag{1}$$

where $\mathbf{X}$ contains m calibration samples measured at n wavelengths, $\mathbf{b}$ is an $n \times 1$ model vector, and $\mathbf{y}$ is an $m \times 1$ vector containing the analytical information for the analyte, such as concentration. To predict future samples, Eq. 1 must first be solved for $\mathbf{b}$, typically accomplished mathematically by $\hat{\mathbf{b}} = \mathbf{X}^{+}\mathbf{y}$, where $\mathbf{X}^{+}$ is a generalized inverse of $\mathbf{X}$. There are several approaches to calculating $\mathbf{X}^{+}$, the most common of which are ridge regression (RR), partial least squares (PLS), and principle component regression (PCR).[1–3] The method of RR is actually a form of Tikhonov regularization (TR)[4,5] as further described in the Mathematics section. With an estimate of $\mathbf{b}$, a new sample spectrum is predicted for its analyte concentration by $\hat{y} = \mathbf{x}^{t}\hat{\mathbf{b}}$, where $\mathbf{x}$ denotes a column vector of the spectral responses.

While the methods of RR, PLS, and PCR can be used with all wavelengths measured (full spectral methods where $n > m$), reduced prediction errors are common when care is taken to select wavelengths spanning useful analyte predictive information.[6–23] However, this prediction error reduction comes with a tradeoff of potential prediction variance inflation.[15,18,22,24,25] The methods of RR, PLS, or PCR can be used when wavelengths are selected such that $n \geq m$ or $n < m$. When the latter is true, the method of multiple linear regression (MLR) can also be used.[1–3]

Methods of wavelength selection are usually one of two modes. One is to select individual wavelengths and the other is to determine wavelength intervals (bands).[13,21,23] Wavelength selection algorithms such as genetic algorithms or simulated annealing commonly necessitate lengthy iterative sequential processes involving wavelength selection, model forming using RR, PLS, or another method, and prediction testing of the selected wavelengths.[6,14–16] These and other optimization algorithms often require user-set operating parameters. Altering these parameters for a specific algorithm can result in different subsets of wavelengths being selected and, hence, the dilemma of selecting wavelengths with chance correlations.[13,17–19] The likelihood of this dilemma occurring increases as the ratio of wavelengths to samples increases.[20]

A common approach to wavelength selection is to rank wavelengths according to a merit that reflects importance and then using an empirically determined cutoff to retain the final set.[6–8,21,26,27] One measure for ranking is the magnitude of coefficients in a full-wavelength model vector, such as would be obtained from PLS. In this case, multiple bands of wavelengths are commonly selected, i.e., adjacent wavelengths often obtain nearly equivalent model coefficients. References 6, 21, 23, and 28 recently reviewed many of these methods as well as interval PLS.[29]

An alternative to wavelength selection algorithms are variants of TR with built-in wavelength selection processes. These variants of TR select wavelengths and build models simultaneously where individual wavelengths and/or multiple bands of wavelengths can be selected. Basically, the full-wavelength TR model coefficients can have values at or near zero. These wavelengths have nearly no affect on prediction and are essentially considered non-selected wavelengths.

Because TR models are formed in a systematic process, the L-curve approach is useful to determine the final wavelength-selected model.[3–5,30] The L-curve approach to meta-parameter selection is well documented. Briefly, a measure reflective of prediction variance is plotted against a prediction-bias diagnostic and the best set of wavelengths are those for models in the corner of the resulting L-curve, i.e., models with an acceptable bias/variance tradeoff. Many wavelength selection algorithms only use a prediction bias information criterion to identify acceptable wavelengths, i.e., the prediction-bias diagnostic is used to guide the algorithm in selecting wavelengths. Algorithmic optimization of only a bias diagnostic such as the root mean square error of calibration (RMSEC) or root mean square error of validation (RMSEV) typically results in over-fitting.[13,22,31–33] The TR wavelength-selection variants require determination of a model meta-parameter (tuning parameter) for an acceptable bias/variance tradeoff. This is no different than other biased modeling methods such as PLS or PCR in full-wavelength modes or in wavelength-selection studies requiring a meta-parameter determination to identify the final model.

Received 27 May 2010; accepted 8 September 2010.

* Author to whom correspondence should be sent. E-mail: kalijohn@isu.edu.

0000–0000/10/6412-1388$2.00/0

© 2010 Society for Applied Spectroscopy

Studied in this paper are three variants of TR used to form models with full wavelengths or selected wavelengths (individual wavelengths and/or wavelength bands). The L-curve is used to select the final models. The three TR approaches include (1) a model vector two-norm (L_2) minimization criterion for a full-wavelength model, (2) a model vector L_2 minimization criterion that includes a priori information to weight model coefficients favoring selection of multiple wavelength intervals, and (3) a model vector one-norm (L_1) minimization criterion that may or may not include a priori information to weight model coefficients favoring individual wavelength selection. Simulated spectral data are assessed to verify algorithm operations. Following this, ultraviolet–visible and near-infrared spectral data sets are evaluated.

MATHEMATICS

Ridge Regression. Ridge regression model vectors for Eq. 1 obtained by TR are those that satisfy

$$\min(||\mathbf{Xb} - \mathbf{y}||_2^2 + \lambda^2||\mathbf{b}||_2^2) \tag{2}$$

where $||\bullet||$ indicates the vector norm and the subscript 2 defines the 2-norm (L_2), also termed the Euclidean norm, and λ is the meta-parameter controlling the weight given to the second term relative to the first term. The 2-norm for $\mathbf{b}$ is computed by $||\mathbf{b}||_2 = (\Sigma_{i=1}^{n} b_i^2)^{1/2}$. For Expression 2, RR models are obtained by varying the λ meta-parameter and computing respective models using

$$\hat{\mathbf{b}} = (\mathbf{X}^t\mathbf{X} + \lambda^2\mathbf{I})^{-1}\mathbf{X}^t\mathbf{y} \tag{3}$$

where $\mathbf{I}$ denotes the identity matrix.

The well documented L-curve approach is valuable in determining a value for λ and, hence, the final model.[3–5,30] The L-curve has also been used to determine the number of basis vectors for PLS and PCR,[3,34] variable selection for MLR,[30] and in TR calibration maintenance and transfer studies.[35–37]

To form the RR L-curve and select λ, model vector 2-norms (the greater the magnitude, the greater the likelihood of overfitting and increased uncertainty in the prediction) are plotted against the respective RMSEC values (a bias indicator). In such a plot, an L-shaped curve is formed and the better models reside in the corner region of the L-curve with acceptable tradeoffs in the plotted criteria. Other model diagnostics, such as R^2, can be plotted.[34]

Tikhonov Regularization in L_2 (TR 2). A more general form of Expression 2 found useful for calibration maintenance and transfer[35–37] is

$$\min(||\mathbf{Xb} - \mathbf{y}||_2^2 + \lambda^2||\mathbf{Mb} - \mathbf{y}_M||_2^2) \tag{4}$$

where $\mathbf{M}$ and $\mathbf{y}_M$ represent a set of spectra and analyte values measured under conditions different than the calibration samples in $\mathbf{X}$ and $\mathbf{y}$. When $\mathbf{M} = \mathbf{I}$ and $\mathbf{y}_M = \mathbf{0}$, Expression 4 reduces to RR in Expression 2 and TR is said to be in standard form.[4] A key variant of Expression 4 is

$$\min(||\mathbf{Xb} - \mathbf{y}||_2^2 + \lambda^2||\mathbf{Mb}||_2^2) \tag{5}$$

The $\mathbf{M}$ matrix has been set to derivative operators in Expression 5 for computing a smoothed model vector[38] and $\mathbf{M}$ has also been set to a diagonal matrix with diagonal elements being spectral noise estimates of the respective wavelengths.[38,39] Using spectral noise estimates is a form of wavelength selection as noisy wavelengths obtain model coefficients at or near zero.

This approach of wavelength selection is again studied in this paper for comparison to using a full-wavelength RR determined model vector on the diagonal of $\mathbf{M}$ where the ith diagonal element is $1/|\hat{b}_i|$ for the ith RR model wavelength coefficient. The idea is that using a priori information about expected model coefficient magnitudes and, hence, possibly a measure of importance, a greater emphasis can be put on key wavelengths (bands) and less emphasis on those wavelengths with near-zero or small RR model coefficients; therefore, a better model vector should result. Other studies have found it useful to select wavelengths based on the magnitudes of PLS model coefficients.[7–9,26] Because the spectral noise structure is not always known, using a full-wavelength model vector on the diagonal of $\mathbf{M}$ is more feasible.

Using Expression 5 will henceforth be referred to as *TR 2* in this paper. When $\mathbf{M} = \mathbf{I}$ in Expression 5, TR 2 will be referred to as RR for distinguishing the full-wavelength form of TR. Computation of the models is obtained from the following

$$\hat{\mathbf{b}} = (\mathbf{X}^t\mathbf{X} + \lambda_2\mathbf{M}^t\mathbf{M})^{-1}\mathbf{X}^t\mathbf{y} \tag{6}$$

and the L-curve approach is used to determine λ.

Tikhonov Regularization in L_1 (LASSO TR). Another general form of TR found useful in calibration maintenance and transfer studies is

$$\min(||\mathbf{Xb} - \mathbf{y}||_2^2 + \tau||\mathbf{Mb} - \mathbf{y}_M||_1^1) \tag{7}$$

where the subscript 1 indicates the vector 1-norm (L_1) and τ represents the weight given to the second term.[35,37] The 1-norm for $\mathbf{b}$ is computed by $||\mathbf{b}||_1 = \Sigma_{i=1}^{n} |b_i|$. If only wavelength selection is desired, then Expression 7 is simplified to

$$\min(||\mathbf{Xb} - \mathbf{y}||_2^2 + \tau||\mathbf{Mb}||_1^1) \tag{8}$$

where, as with Expression 5, $\mathbf{M}$ will be set to a diagonal matrix with spectral noise structure estimates or a previous RR model vector ($1/|\hat{b}_i|$). When $\mathbf{M} = \mathbf{I}$, Expression 8 further reduces to the variable selection method known as the least absolute shrinkage and selection operator (LASSO),[2,40] albeit the approach was developed earlier.[41,42] When $\mathbf{M}$ in Expression 8 is set to a diagonal matrix ($\mathbf{M} \neq \mathbf{I}$), the minimization has been termed adaptive LASSO.[2,43] Using Expression 8 in this paper will be referred to as *LASSO TR* whether $\mathbf{M} = \mathbf{I}$ or $\mathbf{M}$ is a diagonal matrix with either the spectral noise structure or $1/|\hat{b}_i|$ from a RR model.

Tikhonov regularization in the LASSO format with $\mathbf{M} = \mathbf{I}$ has been used for wavelength selection and found to produce lower prediction errors than full-wavelength models.[11,12] In Ref. 12, the L-curve approach with the L_1 norm of the model vectors was used instead of the L_2 norm to determine an acceptable model. Such an L-curve will be used in this study. Reported in this study is the first application of LASSO TR with $\mathbf{M} \neq \mathbf{I}$ for the specific intent of wavelength selection.

The least angle regression (LAR) algorithm[44] is one of several algorithms that can be used for LASSO. At each iteration, a non-zero model vector coefficient in $\mathbf{b}$ is added to or removed from the model vector in the previous iteration. The value $\tau = 0$ forms the classical least squares solution (model

vector in the last iteration of the LAR algorithm). In this paper, the LAR algorithm is used to solve Expression 8 with the modification that the columns of **X** are not scaled to mean zero and standard deviation of one (autoscaled) as originally suggested for LAR.[44] It was found that with autoscaled data, inappropriate wavelengths are selected, i.e., wavelengths with the best sensitivity are not selected for the data sets used in this paper as all wavelengths are seen to have equal sensitivity after autoscaling. When $\mathbf{M} = \mathbf{I}$, the LAR algorithm can be directly used as this is LASSO. However, when **M** is a diagonal matrix, then processes described in Refs. 4, 30, and 45 are used to transform the data and Expression 8 is now written as

$$\min(||\bar{\mathbf{X}}\bar{\mathbf{b}} - \bar{\mathbf{y}}||_2^2 + \tau||\bar{\mathbf{b}}||_1^1) \qquad (9)$$

which can now be solved by the LAR algorithm. The bar indicates transformed data and the LAR-estimated model vector $\hat{\bar{\mathbf{b}}}$ must be transformed back to the estimated model vector $\hat{\mathbf{b}}$ that is used to obtain predicted concentrations. While this transformation process is new to using LASSO TR, the transformation has been used with TR 2, PLS, and PCR in order to obtain smooth model vectors using derivative operators for **M**. The transformation has also been applied with penalty weights in **M** based on spectroscopic noise.[38] The transformation process was also recently used with TR written as Expression 7 for calibration maintenance and transfer.[35,37]

EXPERIMENTAL

Apparatus and Algorithms. MatLab 7.8 (The Math Works, Natick, MA) programs for RR, TR 2, and LASSO TR algorithms as described in the Mathematics section were written by the authors. All programs were run on an Intel Core 2 Quad personal computer.

Data Sets. Two simulated data sets are studied to test the algorithms and understand how the algorithms work, i.e., what type of wavelengths are the TR variants selecting (selectivity versus sensitivity). These two simulated data sets have previously been used.[12,38,39] A near-infrared spectroscopic data set also previously studied[12] is evaluated in this paper. Also assessed is a visible spectroscopic inorganic data set. All samples are mean centered relative to the calibration samples in **X** and **y**. As in previous work, the simulated and near-infrared data sets are split by first sorting samples according to concentration magnitudes in **y** and then every other sample was placed into the validation set, with the remaining samples forming the calibration set. The same procedure was followed for the inorganic data set except all replicate spectra for a sample are placed in the validation set and similarly for the calibration set.

Simulated Set 1. The simulated wavelength selection data set used in Refs. 12, 38, and 39 is again examined in this study. The broad-banded spectral data set is based on a Gaussian curve to simulate the pure-component analyte spectrum at unit concentration over 50 wavelength units while using a second Gaussian curve to simulate an overlapping pure-component interferent spectrum at unit concentration (see Fig. 1a). Random concentrations ranging from 0 to 1 over 66 samples are used to form respective mixture spectra from the product of concentrations with pure-component spectra. Random homoscedastic noise from a normal distribution with a zero mean and standard deviation of one was added to spectra at 1% maximum peak height of the respective mixture spectra

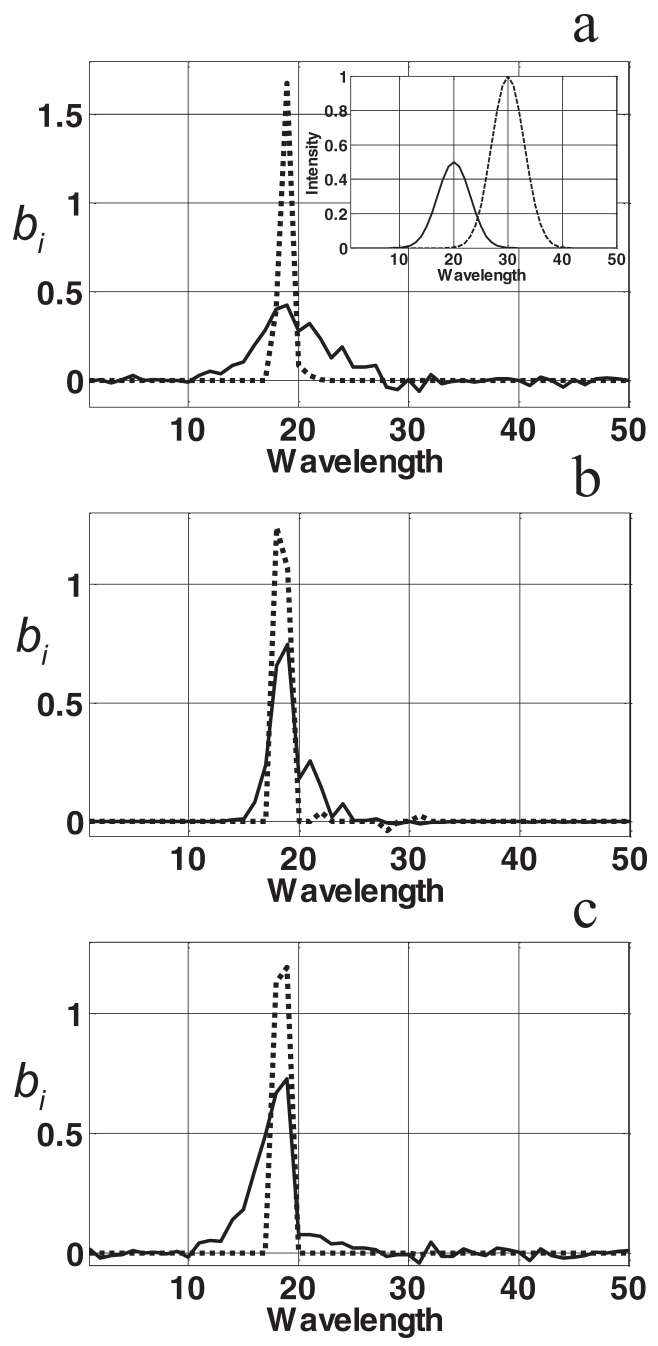

FIG. 1. Simulated set 1 model vectors using RR or TR 2 (solid lines) and LASSO TR (dotted line). (**a**) $\mathbf{M} = \mathbf{I}$ (RR), (**b**) diagonal of **M** set to the RR model vector (TR 2), and (**c**) diagonal of **M** set to the spectral noise structure (TR 2). Pure-component spectra at unit concentration without noise added for the anlayte (solid line) and the interferent (dash line) shown in (**a**).

followed by additional noise at 3% maximum peak height to mixture spectra from wavelengths 20 through 30 (the spectral overlap region).

Simulated Set 2. The simulated wavelength selection data set from Ref. 12 is also reexamined in this study. Narrow-peaked pure-component spectra at unit concentration were created using Gaussian peaks five wavelength units wide with a one-wavelength baseline separating each peak across 61 wavelengths (see Fig. 2a). Analyte peaks are centered at every sixth wavelength from 4 to 52. Interferent peaks are centered at

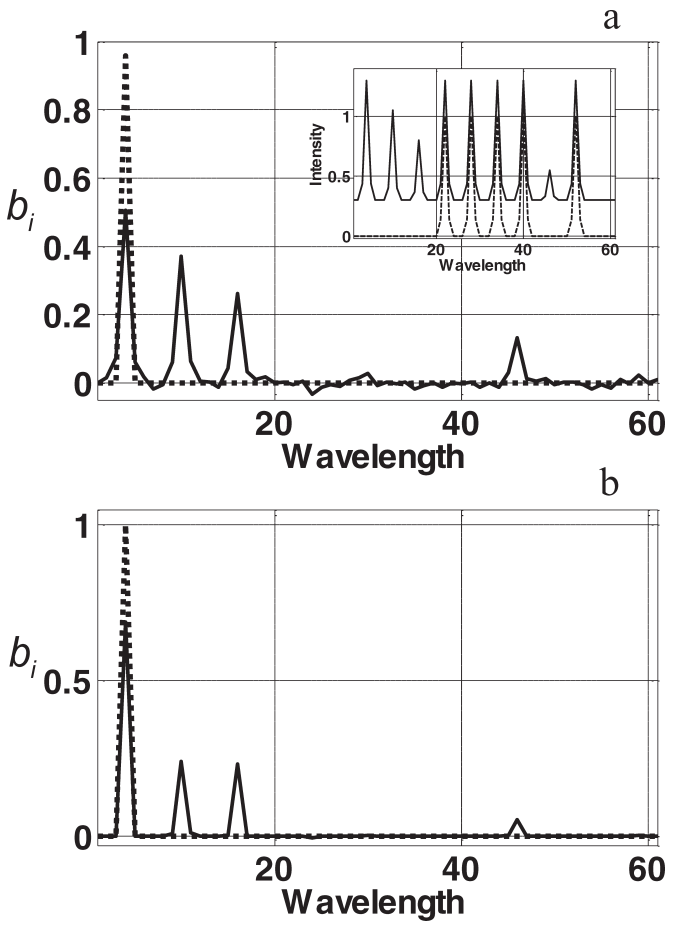

FIG. 2. Simulated set 2 model vectors using RR or TR 2 (solid lines) and LASSO TR (dotted line). (**a**) $\mathbf{M} = \mathbf{I}$ (RR) and (**b**) diagonal of $\mathbf{M}$ set to the RR model vector (TR 2). Pure-component spectra at unit concentration without noise added for the anlayte (solid line) and the interferent (dash line) in (**a**). An offset of 0.3 has been added to the analyte for visual clarity.

wavelengths 22, 28, 34, 40, and 52, producing four peaks with perfect selectivity for the analyte, centered at wavelengths 4, 10, 16, and 46. Sixty-six sample concentrations for both analyte and interferent were created from the absolute values of random numbers with a normal distribution of a mean of zero and a standard deviation of one. Mixture spectra are the product of the concentrations with the pure-component spectra. Random heteroscedastic noise with a normal distribution of mean zero and standard deviation of one was added to the mixture spectra at 1% of respective wavelength spectral values.

Inorganic. The inorganic data set is a three-component system of Co II, Cr III, and Ni II described in Ref. 46. Based on a three-level, three-factor calibration design, 26 samples were prepared. Five replicate spectra were obtained for each sample in randomized blocks creating 128 spectra after removing two spectra as outliers due to an offset. Absorbances were measured from 350 to 650 nm at 2-nm intervals with a diode array spectrophotometer, yielding spectra with 151 wavelengths. Pure-component and mixture spectra are shown in Fig. 3. The standard deviation at each wavelength for each spectrum is also provided. For the purposes of this study Cr was analyzed. The noise structure for $\mathbf{M}$ is the mean of the provided calibration standard deviations.

Corn. The corn data set as used in Ref. 12 is reexamined in this study and consists of 80 corn samples with reference moisture values.[47] Spectra were measured from 1100 to 2498 nm at 2-nm intervals on a near-infrared spectrometer. Every tenth recorded wavelength was used, yielding 70 wavelengths.

RESULTS AND DISCUSSION

Simulated Set 1. When using $\mathbf{M} = \mathbf{I}$, values tabulated in Table I show that LASSO TR produces a lower RMSEV than RR. This comes with a tradeoff with a potential increase in prediction variance as indicted by the increase in the model 2-norm of the regression vector, which agrees with previous work on this data set.[38,39]

When coefficients from a full-wavelength RR model are used as $1/|\hat{b}_i|$ for the diagonal elements of $\mathbf{M}$, results presented in Table I reveal that TR 2 reduces the prediction error compared to using $\mathbf{M} = \mathbf{I}$ for RR. Additionally, prediction results from TR 2 are improved over LASSO TR with $\mathbf{M} = \mathbf{I}$ or the diagonal of $\mathbf{M}$ set to the same full-wavelength RR model. Compared to the RR model vector shown in Fig. 1a, the model from TR 2 plotted in Fig. 1b based on the diagonal of $\mathbf{M}$ set to the RR model vector is more similar to the LASSO TR model vectors plotted in Figs. 1a and 1b with $\mathbf{M} = \mathbf{I}$ or the diagonal of $\mathbf{M}$ set to the RR model vector, respectively. Thus, TR 2 is able to eliminate unnecessary wavelengths, creating a model vector similar to the wavelength selected LASSO TR models. Figures 1a and 1b also show that there is little difference in the LASSO TR model vectors with $\mathbf{M}$ formed as the identity matrix or as the diagonal matrix.

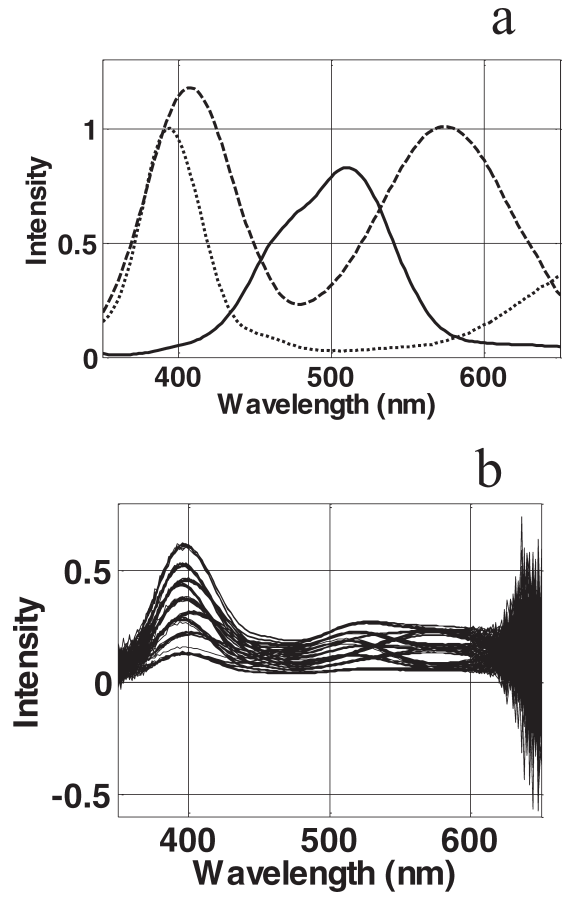

FIG. 3. (**a**) Pure-component spectra of Co (solid line), Cr (dashed line), and Ni (dotted line) and (**b**) mixture spectra.

TABLE I. Simulated data set 1 results.

Method	**M**	RMSEV	R^2	$\|\|\hat{\mathbf{b}}\|\|_2$	λ or τ
RR	**I**	0.0176	0.9951	0.894	0.193
TR 2	RR model	0.0117	0.9984	1.084	0.0468
TR 2	Noise	0.0101	0.9987	1.255	8.08
LASSO TR	**I**	0.0136	0.9977	1.727	0.0145
LASSO TR	RR model	0.0125	0.9985	1.636	3.15×10^{-4}
LASSO TR	noise	0.0126	0.9985	1.643	2.80×10^{-3}

TABLE II. Simulated data set 2 results.

Method	**M**	RMSEV	R^2	$\|\|\hat{\mathbf{b}}\|\|_2$	λ or τ
RR	**I**	0.0141	0.9995	0.711	0.276
TR 2	RR model	0.0131	0.9995	0.766	0.270
LASSO TR	**I**	0.0284	0.9991	0.959	0.469
LASSO TR	RR model	0.0179	0.9992	1.000	2.30×10^{-3}

TABLE III. Inorganic data set Cr results.

Method	**M**	RMSEV	R^2	$\|\|\hat{\mathbf{b}}\|\|_2$	λ or τ
RR	**I**	1.490×10^{-4}	0.9991	0.0158	0.112
TR 2	RR model	1.280×10^{-4}	0.9995	0.0184	8.45×10^{-4}
TR 2	Noise	1.048×10^{-4}	0.9995	0.0175	65.8
LASSO TR	**I**	1.738×10^{-4}	0.9987	0.129	1.63×10^{-4}
LASSO TR	RR model	1.523×10^{-4}	0.9849	0.105	1.42×10^{-7}
LASSO TR	Noise	1.212×10^{-4}	0.9994	0.0860	0.0141

As noted in the Experimental section, the spectral noise structure is known and was used as a diagonal for TR 2 and LASSO TR. Table I shows the results and discloses improvement in the predictive ability for TR 2 and little to no improvement for LASSO TR. From model vectors plotted in Fig. 1c, it is ascertained that TR 2 with the noise structure in **M** generates a model similar to those obtained from LASSO TR in Figs. 1a, 1b, and 1c except more of the selective less sensitive wavelengths are now included in the TR 2 model vector. The inclusion of these wavelengths is probably why this TR 2 model provides the lowest prediction errors.

From the RR model vector in Fig. 1a, wavelengths relative to the analyte are non-zero, including the overlap region with excessive noise. In Figs. 1a, 1b, and 1c for LASSO TR, model vectors are more restrictive and focus on the most sensitive and selective wavelengths. Thus, LASSO TR behaves more as an individual-wavelength selection approach. Intermediate to RR and LASSO TR is TR 2 with a priori information of the full-wavelength RR model (Fig. 1b) or spectral noise (Fig. 1c). In these cases, even more so in Fig. 1c, a band of wavelengths is selected that primarily includes analyte selective wavelengths regardless of sensitivity.

Simulated Set 2. Results tabulated in Table II show that RR performs better than LASSO TR when $\mathbf{M} = \mathbf{I}$. As with simulated set 1, prediction errors improve when the RR model is used as the diagonal of **M** for TR 2 and LASSO TR. Model vectors plotted in Fig. 2 show that compared to the RR model in Fig. 2a, TR 2 in Fig. 2b with the RR model on the diagonal of **M** reduces the model coefficients in the spectral interferent region noted in Fig. 2a. For RR and TR 2, the focus is on the selective wavelengths regardless of sensitivity.

The LASSO TR model vectors plotted in Fig. 2 reveal that only the most sensitive and selective wavelength at position 4 is included. The only difference between using $\mathbf{M} = \mathbf{I}$ and the RR model on the diagonal of **M** is the slight increase of the model coefficient at wavelength 4. As is shown in Table II, all improvements in prediction errors come with a sacrifice in potential increases in prediction variance as indicated by the increased model 2-norms.

Inorganic. Results tabulated in Table III show similar trends for the two simulated data sets. Specifically, using a priori information of the spectral noise structure or a RR model vector on the diagonal of **M** provides improved prediction results for TR 2 and LASSO TR compared to using $\mathbf{M} = \mathbf{I}$. From model vectors plotted in Figs. 4a and 4b, it is observed that RR and TR 2 model vectors are similar in shape. When the spectral noise structure is included in **M** for TR 2, the noisy wavelength regions are now zeroed out (Fig. 4c). The LASSO TR model vectors plotted in Fig. 4 reveal little differences in the three models for $\mathbf{M} = \mathbf{I}$, **M** with a RR model on the diagonal, and **M** with the spectral noise structure. The LASSO TR models tend to have selected individual wavelengths and are outperformed in terms of prediction error by TR 2 with selection of wavelength bands.

Corn. Results listed in Table IV show improvement in predictions when using LASSO TR over RR with $\mathbf{M} = \mathbf{I}$. Further respective prediction improvements occur when the diagonal of **M** is set to a RR model vector. Model vectors pictured in Fig. 5 are characterized by comparable trends observed for the other data sets. A notable difference is LASSO TR selecting bands of wavelengths, whereas in the previous data sets the focus is on individual wavelengths. The TR 2 and

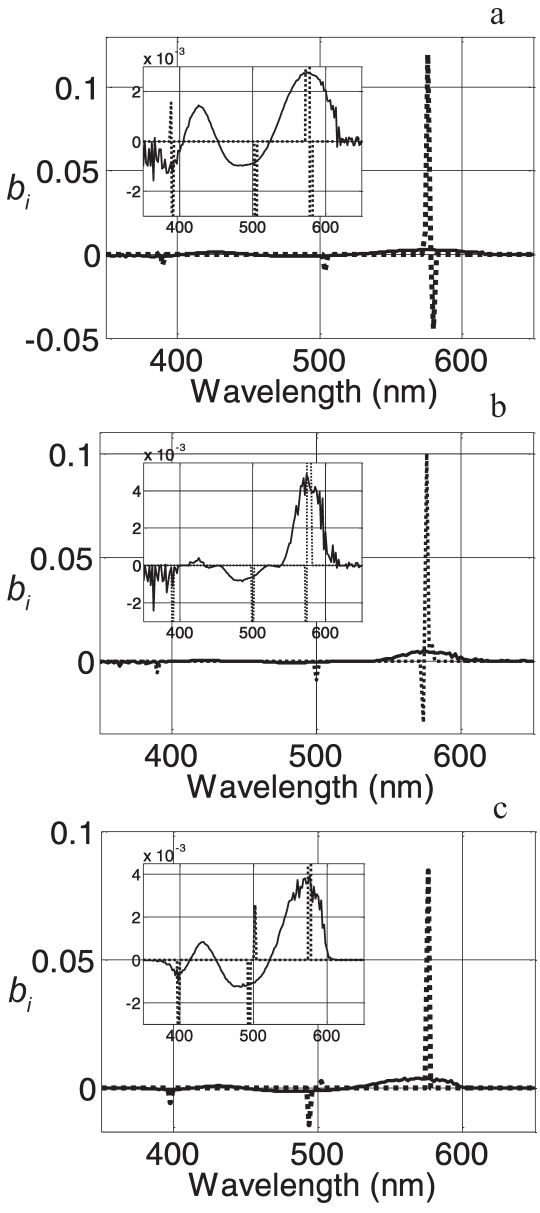

FIG. 4. Inorganic Cr model vectors using RR or TR 2 (solid lines) and LASSO TR (dotted line). (**a**) $\mathbf{M} = \mathbf{I}$ (RR), (**b**) diagonal of **M** set to the RR model vector (TR 2), and (**c**) diagonal of **M** set to the spectral noise structure (TR 2).

LASSO TR model vectors are similar to those shown in Ref. 12 obtained by using a different LASSO TR algorithm. These model vectors are also nearly the same as those recently reported in Ref. 8 that selects and weights wavelengths based on PLS model coefficient magnitudes.

Simultaneous TR 2 and LASSO TR. A more general form of TR that incorporates Expressions 4 and 7 as special cases and, hence, also contains RR, LASSO, and adaptive LASSO as special cases is

$$\min(||\mathbf{Xb} - \mathbf{y}||_2^2 + \lambda^2||\mathbf{Mb} - \mathbf{y}_M||_2^2 + \tau||\mathbf{Eb} - \mathbf{y}_E||_1^1) \quad (10)$$

where **M** and **E** can or cannot be the same matrix. A variant of Expression 10 has been used for calibration maintenance and transfer and has been shown to have significant advantages with regard to robustness of the sample composition in **M**.[35] For only wavelength selection, Expression 10 simplifies to

$$\min(||\mathbf{Xb} - \mathbf{y}||_2^2 + \lambda^2||\mathbf{Mb}||_2^2 + \tau||\mathbf{Eb}||_1^1) \quad (11)$$

where **M** and **E** are diagonal matrices set to the spectral noise structure, a previous model vector, the identity matrix, or another weighting matrix. In the special case where **M** and **E** are both set to **I**, then Expression 11 has been termed the elastic net.[2,48] Similar to Expression 9, the data is transformed to allow using the LAR algorithm. Before transformation, calibration spectra and concentrations are first augmented with $\lambda\mathbf{M}$ and **0** to form

$$\mathbf{X}_A = \begin{pmatrix} \mathbf{X} \\ \lambda\mathbf{M} \end{pmatrix}$$

and

$$\mathbf{y}_A = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

Expression 11 is now written as

$$\min(||\mathbf{X}_A\mathbf{b} - \mathbf{y}_A||_2^2 + \tau||\mathbf{Eb}||_1^1) \quad (12)$$

Data in Expression 12 are then transformed as with Expression 8 to be in the format of Expression 9 and the LAR algorithm can now be used. This augmentation and transformation process has been used in other studies.[35,37]

In using Expression 11, the L_2 criterion weights the minimization towards a full-wavelength model in conjunction with an L_1 criterion weighting the minimization towards a model with wavelengths selected. Studies of this nature were undertaken and it was found that the models formed prediction errors intermediate to TR 2 and LASSO TR depending on the values for tuning parameters λ and τ in Expression 11. For a fixed τ value and λ increasing, the model converges to TR 2; vice-versa, for a fixed λ value and τ increasing, the model converges to LASSO TR. These observations hold true for all the data sets examined in this study regardless of the structure of **M** and **E**. Thus, this TR variant was not found useful as either TR 2 or LASSO TR always outperformed it. However, this variant has been found to be effective in calibration maintenance and transfer work.[35,37]

TABLE IV. Corn data set moisture results.

Method	**M**	RMSEV	R^2	$\|\|\hat{\mathbf{b}}\|\|_2$	λ or τ
RR	**I**	0.0227	0.9972	91.2	5.88×10^{-4}
TR 2	RR model	0.0130	0.9992	106.4	8.70×10^{-3}
LASSO TR	**I**	0.0156	0.9988	113.7	2.27×10^{-5}
LASSO TR	RR model	0.0113	0.9994	114.7	6.25×10^{-5}

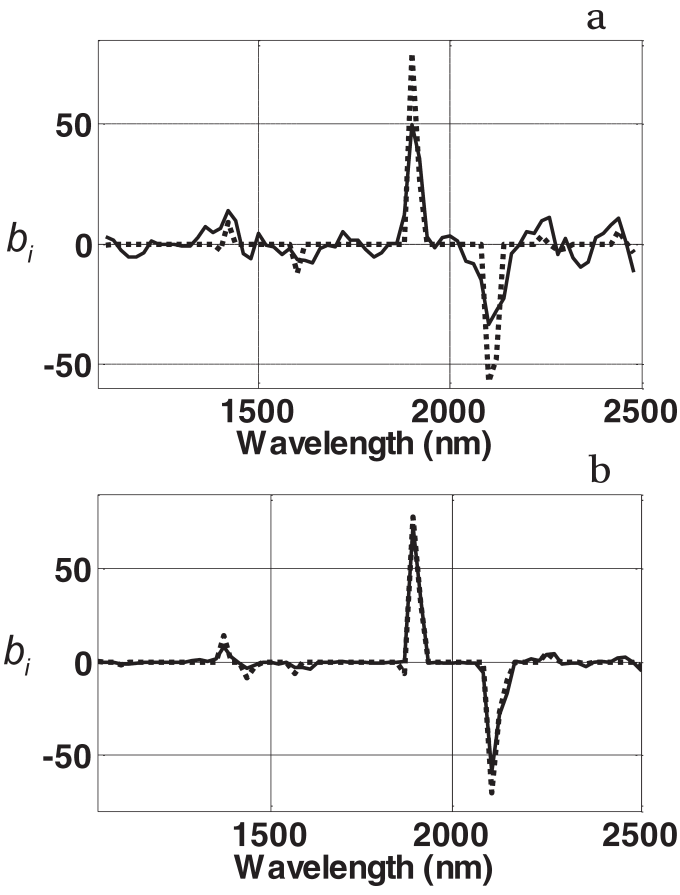

FIG. 5. Corn moisture model vectors using RR or TR 2 (solid lines) and LASSO TR (dotted line). (**a**) $\mathbf{M} = \mathbf{I}$ (RR) and (**b**) diagonal of **M** set to the RR model vector (TR 2).

CONCLUSION

For all the data sets, some type of wavelength selection was found to provide reduced predictions errors compared to the full-wavelength models from RR. Generally, TR 2 with a RR model on the diagonal of **M** formed models with bands of wavelengths selected and lower prediction errors than the LASSO TR models. When the spectral noise structure is known, then TR 2 with the noise on the diagonal of **M** proved to provide the lowest prediction errors. While no specific significance tests were applied to the RMSEV values from the various approaches, the similar trends between data sets indicate that the differences are relevant. The full-wavelength RR model vector was used for the diagonal of **M** in this study; however, other full-wavelength models could be used such as from PLS. For the data sets evaluated, the analyte is present at moderate levels and similar results would be expected for situations with the analyte at reduced levels.

The focus of the paper is on wavelength selection, but the TR variants are generic and can be applied to other data sets in need of variable selection. Additionally, the presented variants of TR could be applied iteratively. For example, the model resulting from TR 2 with the diagonal of **M** set to a full-wavelength RR model vector could then be used as the diagonal of **M** for another TR 2 and so on until convergence of the model vector. A similar iterative process could be used with LASSO TR.

An interesting variant of LASSO TR with $\mathbf{M} = \mathbf{I}$ is constraining model coefficients to lie within a suitable polyhedral region.[49] A first guess of the polyhedral could be based on coefficient confidence intervals for a preliminary model vector obtained by RR or another approach. A method not tested is to use respective estimated coefficient uncertainties for a full-wavelength model vector as the diagonal of **M**, e.g., regression vector coefficient uncertainties for a full-wavelength RR model.

Another variant of TR is fused LASSO,[50] in which two L_1 minimizations are included. One is for the model vector to obtain variable selection and the other is for piece-wise consistency (variable intervals with constant model coefficient values). Because the second penalty forces a band of wavelengths to have constant model coefficient values, the fused LASSO approach is not expected to provide reduced prediction errors. However, it may be useful in identifying wavelength bands and then only these bands are used in RR or another method. Such an approach was not attempted.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CHE-0715179 (co-funded by the MPS Chemistry and DMS Statistics Divisions, and by the MSPA Program) and is gratefully acknowledged by the authors.

1. T. Næs, T. Isaksson, T. Fern, and T. Davies, *A User Friendly Guide to Multivariate Calibration and Classification* (NIR Publications, Chichester, 2002).
2. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, 2001).
3. J. H. Kalivas, "Calibration Methodologies", in *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, S. D. Brown, R. Tauler, B. Walczak, Eds.-in-Chief, and J. H. Kalivas, Section Ed. (Elsevier, Amsterdam, 2009), Vol. 3, Chap. 1, pp. 1–32.
4. P. C. Hansen, *Rank Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion* (SIAM, Pennsylvania, 1998).
5. R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems* (Elsevier, Amsterdam, 2005).
6. R. K. H. Galvão and M. C. U. Araújo, "Variable Selection", in *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, S. D. Brown, R. Tauler, B. Walczak, Eds.-in-Chief, and J. H. Kalivas, Section Ed. (Elsevier, Amsterdam, 2009), Vol. 3, Chap. 5, pp. 233–283.
7. R. F. Teófilo, J. P. A. Martins, and M. M. C. Ferreira, J. Chemom. **23,** 32 (2009).
8. H. Li, Y. Liang, Q. Xu, and D. Cao, Anal. Chim. Acta **648,** 77 (2009).
9. U. Norinder, J. Chemom. **10,** 95 (1996).
10. H. Öjelund, P. J. Brown, H. Madsen, and P. Thyregod, Technometrics **44,** 369 (2002).
11. I. Chong and C. Jun, Chemom. Intell. Lab. Syst. **78,** 103 (2005).
12. F. Stout, J. H. Kalivas, and K. Héberger, Appl. Spectrosc. **61,** 85 (2007).
13. J. M. Brenchley, U. Hörchner, and J. H. Kalivas, Appl. Spectrosc. **51,** 689 (1997).
14. P. J. de Groot, H. Swierenga, G. J. Postma, W. J. Melssen, and L. M. C. Buydens, Appl. Spectrosc. **57,** 642 (2003).
15. J. Jiang, R. J. Berry, H. W. Siesler, and Y. Ozaki, Anal. Chem. **74,** 3555 (2003).
16. M. L. Griffiths, D. Svozil, P. Worsfold, S. Denham, and E. H. J. Evans, J. Anal. At. Spectrom. **17,** 800 (2002).
17. H. Mark, Appl. Spectrosc. **42,** 1427 (1988).
18. H. Mark, *Principles and Practice of Spectroscopic Calibration* (Wiley, New York, 1991).
19. J. G. Topliss and R. P. Edwards, J. Medicinal Chem. **22,** 1238 (1979).
20. R. Leardi and A. L. Gonzalez, Chemom. Intell. Lab. Syst. **41,** 195 (1998).
21. R. Gosselin, D. Rodrigue, and C. Duchesne, Chemom. Intell. Lab. Syst. **100,** 12 (2009).
22. K. J. Anderson and J. H. Kalivas, Appl. Spectrosc. **57,** 309 (2003).
23. Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, and M. Hanpin, Anal. Chim. Acta **667,** 14 (2010).
24. A. Lorber and B. R. Kowalski, J. Chemom. **2,** 93 (1988).

25. S. D. Frans and J. M. Harris, Anal. Chem. **57,** 2680 (1985).
26. K. Wongravee, N. Heinrich, M. Holmboe, M. L. Schaefer, R. R. Reed, J. Trevjo, and R. G. Brereton, Anal. Chem. **81,** 5204 (2009).
27. W. Wu, Q. Guo, D. Joun-Rimbaud, and D. L. Massart, Chemom. Intell. Lab. Syst. **45,** 39 (1999).
28. N. Sorol, E. Arancibia, S. A. Bortolato, and A. C. Olivieri, Chemom. Intell. Lab. Syst. **102,** 100 (2010).
29. A. S. L. Nøgaard, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, Appl. Spectrosc. **54,** 413 (2000).
30. C. L. Lawson and R. J. Hanson, *Solving Least Square Problems* (SIAM, Pennsylvania, 1995).
31. P. J. Brown, J. Chemom. **6,** 151 (1992).
32. G. W. Small, M. A. Arnold, and L. A. Marquardt, Anal. Chem. **65,** 3279 (1993).
33. R. E. Schaffer, G. W. Small, and M. A. Arnold, Anal. Chem. **68,** 2663 (1996).
34. J. B. Forrester and J. H. Kalivas, J. Chemom. **18,** 372 (2004).
35. M. R. Kunz, J. Ottaway, J. H. Kalivas, and E. Andries, J. Chemom. **24,** 218 (2010).
36. J. H. Kalivas, G. G. Siano, E. Andries, and H. C. Goicoechea, Appl. Spectrosc. **63,** 800 (2009).
37. M. R. Kunz, J. H. Kalivas, and E. Andries, Anal. Chem. **82,** 3642 (2010).
38. F. Stout and J. H. Kalivas, J. Chemom. **20,** 22 (2006).
39. J. H. Kalivas, Anal. Chim. Acta **505,** 9 (2004).
40. R. Tibshirani, J. R. Stat. Soc. B. **58,** 267 (1996).
41. J. F. Claerbout and F. Muir, Geophysics. **38,** 826 (1973).
42. H. L. Taylor, S. C. Banks, and J. F. McCoy, Geophysics **44,** 39 (1979).
43. H. Zou, J. Am. Stat. Assoc. **101,** 1418 (2006).
44. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, R. Annals Statist. **32,** 407 (2004).
45. L. Elden, BIT **27,** 487 (1982).
46. P. D. Wentzel, D. T. Andrews, and B. R. Kowalski, Anal. Chem. **69,** 2299 (1997).
47. H. Zou and T. Hastie, J. R. Stat. Soc. B. **67,** 301 (2005).
48. B. M. Wise, N. B. Gallagher, R. Bro, and J. M. Shaver, *PLS_Toolbox 3.0 for Use with MATLAB* (Eigenvector Research, Washington, 2003).
49. B. A. Turlach, W. N. Venables, and S. J. Wright, Technometrics **47,** 349 (2005).
50. R. Tibshirani, M. Saunders, J. Zhu, and K. Knight, J. R. Stat. Soc. B **67,** 91 (2005).