focal point review

ERIK ANDRIES*
CENTER FOR ADVANCED RESEARCH COMPUTING,
UNIVERSITY OF NEW MEXICO, ALBUQUERQUE,
NEW MEXICO, 87106 USA

SHAWN MARTIN
DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF OTAGO,
P.O. BOX 56, DUNEDIN 9054, NEW ZEALAND

Sparse Methods in Spectroscopy: An Introduction, Overview, and Perspective

Multivariate calibration methods such as partial least-squares build calibration models that are not parsimonious: all variables (either wavelengths or samples) are used to define a calibration model. In high-dimensional or large sample size settings, interpretable analysis aims to reduce model complexity by finding a small subset of variables that significantly influences the model. The term "sparsity", as used here, refers to calibration models having many zero-valued regression coefficients. Only the variables associated with non-zero coefficients influence the model. In this paper, we briefly review the regression problems associated with sparse models and discuss their spectroscopic applications. We also discuss how one can re-appropriate sparse modeling algorithms that perform wavelength selection for purposes of sample selection. In particular, we highlight specific sparse modeling algorithms that are easy to use and understand for the spectroscopist, as opposed to the overly complex "black-box" algorithms that dominate much of the statistical learning literature. We apply these sparse modeling approaches to three spectroscopic data sets.

Received 29 January 2013; accepted 8 March 2013.
* Author to whom correspondence should be sent. E-mail: erik.andries@gmail.com.
DOI: 10.1366/13-07021

Index Headings: **Sparsity; Wavelength selection; Sample selection; Partial least square; Support-vector regression.**

INTRODUCTION

A disadvantage of traditional multivariate calibration (MC) methods such as partial least squares (PLS) and principal component regression (PCR) is that they produce calibration models that are not parsimonious; all of the regression coefficients in the calibration model (or regression vector) are non-zero. As a result, all wavelengths are used in the prediction of unknown samples. Such methods offer little insight into the relative physical importance of different wavelengths. In addition, the weight or influence from irrelevant wavelengths, however small, can have deleterious effect on prediction, since they contribute to unwanted noise.

In spectroscopy, there is a need to prune large data sets, in both the number of samples and wavelengths, to a manageable size, to find only those samples and wavelengths that meaningfully span useful analyte-predictive information. ''Wavelength selection'' refers to the identification of wavelengths relevant to a particular calibration model. In spectroscopy, wavelength selection has a long history; notable examples include interval PLS,[1] genetic algorithms,[2,3] selectivity ratios,[4] and variables important for projection.[5] However, in the last decade, advances in sparse methods, a new class of computationally efficient wavelength-selection methods, have made the problem of reducing model complexity tractable for large data sets.[6–12] The intent of this paper is to give an introduction, overview, and perspective on sparse methods.

Our perspective on sparse methods focuses on how wavelength and sample selection can be performed with the same core set of algorithms. Moreover, for the spectroscopist and the lay statistician, these algorithms need not be overly complex to understand and implement. According to this perspective, we distinguish between two types of sparse methods, sparse methods for wavelength selection and sparse methods for sample selection. In the case of

wavelength selection, a regression vector is generated, such that each regression coefficient is associated with a particular wavelength. The larger the coefficient magnitude, the greater the weight or influence associated with the corresponding wavelength. Similarly, with respect to sample selection, a regression vector is generated, such that each coefficient corresponds to a particular sample.

The paper is organized as follows: We first describe sparse methods in general for wavelength selection and discuss some of their spectroscopic applications. Next, specific approaches and implementations for wavelength selections by using sparse methods are reviewed. Then, we outline how sparse methods for wavelength selection can be repurposed for sample selection. Regression examples from three spectroscopic data sets are examined with sparse methods, followed by the conclusion and discussion of future work.

As for notation in this paper, lowercase and uppercase letters that are not boldfaced correspond to scalars (e.g., x or P). Lowercase and uppercase boldface symbols represent column vectors (e.g., $\mathbf{x}$) and matrices (e.g., $\mathbf{X}$). The subscripted symbols ^{T}, $^{-1}$, and $^{+}$ indicate the transpose, inverse, and pseudoinverse, respectively. The symbol $\mathbf{I}_n$ represents the identity matrix of dimension n. A diagonal matrix is indicated via the "diag" notation, e.g., $\mathbf{I}_n = \text{diag}(1, 1, \ldots, 1)$.

A vector of n ones or n zeros is indicated by $\mathbf{1}_n$ or $\mathbf{0}_n$, respectively. The element associated with the ith row and jth column of the matrix $\mathbf{A}$ will be denoted in two ways, a_{ij} or A_{ij}.

In this paper, the $m \times n$ matrix $\mathbf{X}$ represents spectra measured across m samples and across n wavelengths or frequencies, in which $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m]^T$, where $\mathbf{x}_j = [\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn}]^T$. The vector $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m]^T$ contains the analyte concentration for each sample. The aim of multivariate calibration is to estimate a model or regression vector $\mathbf{b} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n]^T$ that relates $\mathbf{X}$ to $\mathbf{y}$ with best accuracy and precision. We use the notation $\mathbf{b}^{[i]}$ to denote the ith regression vector obtained in an iterative scheme. For the vector $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, we use the following two vector norms to indicate its length or size, the one norm $||\mathbf{x}||_1 = |x_1| + \ldots + |x_n|$ and the two norm $||\mathbf{x}||_2 = \sqrt{x_1^2 + \ldots + x_n^2}$. It is commonplace to mean center the calibration data $\mathbf{X} = \mathbf{X} - 1_m\bar{\mathbf{x}}^T$ and $\mathbf{y} = \mathbf{y} - 1_m\bar{\mathbf{y}}$, in which $\bar{\mathrm{x}}$ and $\bar{\mathrm{y}}$ denote the mean spectrum and mean response, respectively, of the calibration samples. Prediction on an unseen spectrum $\mathbf{z}$ is then given by $f(\mathbf{z}) = (\mathbf{z} - \bar{\mathbf{x}})^T\mathbf{b} + \bar{\mathbf{y}}$.

THEORY: SPARSE METHODS FOR WAVELENGTH SELECTION

Most spectral data sets are characterized by a high-dimensional, low sample-size setting, i.e., there are many more wavelengths than samples. The number n of wavelengths is typically on the order of hundreds, thousands, or higher. As a result, there is a strong desire to separate meaningful wavelength features from spurious or noisy ones. Sparse methods provide a vehicle for performing this separation.

$$b = (S(\lambda) + c)/a$$

Standard Formulations. The minimization problem $\min_{\mathbf{b}} \frac{1}{2}||\mathbf{Xb} - \mathbf{y}||_2^2$ associated with ordinary least squares generally overfits the data, a high-variance (precision), low-bias (accuracy) scenario. If we sacrifice some bias in favor of decreased variance, then overall prediction accuracy might be improved. One way to decrease the variance and increase the bias is to shrink the size of $\mathbf{b}$ by adding a penalty term $P(\mathbf{b})$ to $\frac{1}{2}||\mathbf{Xb} - \mathbf{y}||_2^2$. Different penalties generate different small-norm solutions. The two-norm penalty $P(\mathbf{b}) = \frac{1}{2}\eta||\mathbf{b}||_2^2$ corresponds to ridge regression[13] or Tikhonov regularization (TR):[14]

$$\min_{\mathbf{b}} \frac{1}{2}\left\|\mathbf{Xb} - \mathbf{y}\right\|_2^2 + \frac{1}{2}\eta\left\|\mathbf{b}\right\|_2^2 \qquad (1)$$

Although η controls the size of $\mathbf{b}$ (larger η yields smaller-norm regression vectors), none of the regression coefficients b_i is fully suppressed to zero.

The two-norm penalty in Eq. 1 can be replaced with the one-norm penalty $P(\mathbf{b}) = \lambda||\mathbf{b}||_1$:

$$\min_{\mathbf{b}} \frac{1}{2}\left\|\mathbf{Xb} - \mathbf{y}\right\|_2^2 + \lambda\left\|\mathbf{b}\right\|_1 \qquad (2)$$

The one-norm penalty shrinks many of the coefficients b_i to zero. (See the "Shooting Algorithm" section below for an intuitive discussion of how the one-norm penalty "zeros out" coefficients.) As a result, only the wavelengths associated with non-zero coefficients play any role in prediction. This, in effect, is wavelength selection. The one-norm penalty parameter λ controls both the size and sparsity of $\mathbf{b}$: A larger λ yields a greater number of zero-valued coefficients.

The general idea of using one-norm penalty methods for feature selection dates back to the 1970s in the geophysics literature.[15–18] Two decades later in statistics, the idea of using one-norm penalty methods in regression was popularized by Robert Tibshirani, who coined the acronym LASSO, which stands for least absolute shrinkage and selection operator.[19] The term LASSO has come to dominant the literature when describing the minimization problem of Eq. 2.

The penalty terms in Eq. 1 and Eq. 2 give each regression coefficient of $\mathbf{b}$ equal weight. These penalty terms can be easily modified to give the coefficients unequal weights:

$$\min_{\mathbf{b}} \frac{1}{2}\left\|\mathbf{Xb} - \mathbf{y}\right\|_2^2 + \frac{1}{2}\eta\left\|\mathbf{Lb}\right\|_2^2 \text{ and}$$
$$\min_{\mathbf{b}} \frac{1}{2}\left\|\mathbf{Xb} - \mathbf{y}\right\|_2^2 + \lambda\left\|\mathbf{Lb}\right\|_1 \qquad (3)$$

The matrix $\mathbf{L}$ in Eq. 3 is diagonal and consists of positive entries. In the statistics literature, the one-norm penalty formulation of Eq. 3 is known as the "adaptive LASSO."[20] By using the following variable transformations $\overline{\mathbf{X}} = \mathbf{XL}^{-1}$ and $\bar{\mathbf{b}} = \mathbf{Lb}$, the minimization problems in Eq. 3 can be rewritten in terms of their standard penalty formulations in Eq. 1 and Eq. 2:

$$\min_{\bar{\mathbf{b}}} \frac{1}{2}\left\|\overline{\mathbf{X}}\bar{\mathbf{b}} - \mathbf{y}\right\|_2^2 + \frac{1}{2}\eta\left\|\bar{\mathbf{b}}\right\|_2^2 \text{ and}$$
$$\min_{\bar{\mathbf{b}}} \frac{1}{2}\left\|\overline{\mathbf{X}}\bar{\mathbf{b}} - \mathbf{y}\right\|_2^2 + \lambda\left\|\bar{\mathbf{b}}\right\|_1 \qquad (4)$$

After solving for $\bar{\mathbf{b}}$ in Eq. 4, the model vector $\mathbf{b}$ of interest can be recovered via the relation $\mathbf{b} = \mathbf{L}^{-1}\bar{\mathbf{b}}$. The variable transformation $\overline{\mathbf{X}} = \mathbf{XL}^{-1}$ can be seen as a rescaling of the spectra $\mathbf{X}$ on a

wavelength-by-wavelength basis. Later, we examine how this rescaling can be exploited for purposes of wavelength selection.

Multiple-Norm Formulations. Numerical problems in Eq. 2 and Eq. 4 can occur when (i) there are fewer samples than wavelengths or (ii) the spectral samples are highly collinear. Either condition renders **X** rank deficient, and as a consequence, numerical instabilities arise. To correct this, we can combine both the two-norm and one-norm penalties of Eq. 1 and Eq. 2, respectively, into a single mixed-norm expression:

$$\min_{\mathbf{b}} \tfrac{1}{2}\left\|\mathbf{X}\mathbf{b}-\mathbf{y}\right\|_2^2+\tfrac{1}{2}\eta\left\|\mathbf{b}\right\|_2^2+\lambda\left\|\mathbf{b}\right\|_1 \tag{5}$$

Eq. 5 can also be rewritten in an augmented fashion:

$$\min_{\mathbf{b}} \tfrac{1}{2}\left\|\mathbf{X}_\eta\mathbf{b}-\mathbf{y}_\eta\right\|_2^2+\lambda\left\|\mathbf{b}\right\|_1,$$
$$\mathbf{X}_\eta=\begin{bmatrix}\mathbf{X}\\ \eta\mathbf{I}_n\end{bmatrix} \text{ and } \mathbf{y}_\eta=\begin{bmatrix}\mathbf{y}\\ \mathbf{0}_n\end{bmatrix} \tag{6}$$

Note that minimizing the first two terms in Eq. 5 or the first term in Eq. 6 is equivalent to TR, the minimization of Eq. 1. Effectively, the two-norm penalty in Eq. 5 forces **X** to be full rank by augmenting it with a multiple of the identity matrix. The phrase "elastic net" often is used to refer to the minimization problem associated with Eq. 5 or Eq. 6.[21]

In addition to overcoming numerical instabilities inherent in the original LASSO formulation of Eq. 2, the elastic net formulation of Eq. 5 also addresses another shortcoming of the LASSO: the selection of a single wavelength versus the selection of an entire interval of highly correlated and neighboring wavelengths. If there is a group or interval of wavelengths among which the pairwise correlations are very high (e.g., neighboring wavelengths associated with a particular type of chemical bonding), then the LASSO tends to select only one wavelength from this group. This selection of a single archetypal wavelength from the group can degrade performance relative to the useful redundancy achieved by using all wavelengths in a group. A detailed discussion of the grouping effect of the elastic net can be found in Zou and Hastie's work.[21]

Tuning Parameters. In the LASSO problem of Eq. 2, one needs to optimize the one-norm penalty parameter λ on the basis of the calibration spectra. If the elastic net variant of the LASSO in Eq. 6 is used, then two parameters must be optimized, λ and η. Cross-validation (CV) is one of the most commonly used mechanisms for determining λ and η, but there are others. Examples include the L-curve,[22,23] Akaike or Bayesian information criteria,[24,25] the F-test,[26] and bootstrapping.[27]

In the case of CV, for example, a diagnostic figure of merit such as the root mean square error of CV (RMSECV) is often plotted for each (λ, η) pair. Naïvely, one could choose the parameter pair with the lowest RMSECV, but this usually results in parameters that overfit the calibration data.[26] There are many other figures of merit that one can use in conjunction with RMSECV. An examination and discussion of spectroscopically relevant figures of merit for various LASSO variants was recently undertaken by Kalivas.[23]

Each penalty parameter has an effective range or interval from which applicable nonnegative values can be sampled. In the case of the two-norm penalty parameter η, the effective range is the interval $[0, \eta_{max}]$, where η_{max} is the largest singular value of **X**.[22] In the case of the one-norm penalty parameter λ, the effective range of values is the interval $[0, \lambda_{max}]$, where $\lambda_{max} = \{|c_1|, \ldots, |c_n|\}$, such that $\mathbf{c} = \mathbf{X}^T\mathbf{y}$.[28] A common sampling strategy for either parameter is to select N values in an exponentially decaying fashion, starting from the maximal value and ending at zero.[22]

APPLICATIONS: SPARSE METHODS FOR WAVELENGTH SELECTION

Although regression techniques such as LASSO achieve parsimonious calibration models, the selection of wavelengths that span useful analyte-predictive information is not the only reason for their utility in spectroscopic analysis. We briefly highlight some additional applications.

Calibration Maintenance and Transfer. A calibration model has limited applicability over time. The primary spectra **X** represent the original state of instrumental, chemical, physical, and/or environmental conditions when the spectra were originally measured. Calibration maintenance seeks to maintain the primary calibration model for spectra measured under new secondary conditions that were not spanned in the original calibration domain. One way to accomplish this task is to augment the secondary spectra to the primary spectra. The concern of calibration transfer, on the other hand, is in using a calibration model developed under a primary instrument to predict spectral samples measured on a secondary instrument. (The secondary instrument can also be the primary instrument at a later point.)

Suppose we replace the penalty $\frac{1}{2}\eta^2||\mathbf{b}||_2^2$ in Eq. 5 with $\frac{1}{2}\eta^2||\mathbf{M}\mathbf{b}||_2^2$, where **M** is a matrix of secondary spectra. In particular, suppose **M** is a $p \times n$ matrix of interferent spectra; spectra not due to analyte variation, but from spectral interferences resulting from any combination of effects stemming from chemical, physical, environmental, or instrumental sources. For example, when detecting an analyte not normally present in a sample (e.g., alcohol in human tissue), the spectra **X** and **M** could be made up of samples having non-zero- and zero-valued analyte concentrations, respectively. When a sample does not (or cannot) have zero-analyte concentration, the matrix of interferent spectra could be obtained from samples having constant analyte concentrations. The mean centering of the constant-analyte spectra would then yield **M**.

By using the matrix of interferent spectra **M** instead of the identity matrix $\mathbf{I}_n$, Eq. 6 can be rewritten as:

$$\min_{\mathbf{b}} \tfrac{1}{2}\left\|\mathbf{X}_\eta\mathbf{b}-\mathbf{y}_\eta\right\|_2^2+\lambda\left\|\mathbf{b}\right\|_1, \mathbf{X}_\eta = \begin{bmatrix}\mathbf{X}\\ \eta\mathbf{M}\end{bmatrix} \text{ and } \mathbf{y}_\eta=\begin{bmatrix}\mathbf{y}\\ \mathbf{0}_p\end{bmatrix} \tag{7}$$

Minimizing $\frac{1}{2}||\mathbf{X}_\eta\mathbf{b}-\mathbf{y}_\eta||_2^2$ by itself in Eq. 7 approximates the following equality-constrained least-squares problem: Solve $\mathbf{X}\mathbf{b} = \mathbf{y}$, subject to $\mathbf{M}\mathbf{b} = \mathbf{0}_p$. Geometrically, the constraints $\mathbf{M}\mathbf{b} = \mathbf{0}_p$ are a statement about orthogonality: as η increases, **b** is increasingly perpendicu-

lar to the space spanned by the spectra in $\mathbf{M}$. Spectroscopically, the constraints $\mathbf{Mb} = \mathbf{0}_p$ attempt to desensitize the primary calibration model against spectral artifacts (the secondary conditions) by pointing $\mathbf{b}$ away from the space spanned by spectral interferents containing no analyte information.

Alternatively, Eq. 7 could also represent a calibration transfer scenario. Suppose $\mathbf{X}_1$ and $\mathbf{X}_2$ are spectra measured from two different instruments (the first and second instruments, respectively). The difference spectra $\mathbf{M} = \mathbf{X}_1 - \mathbf{X}_2$ capture the spectral variation associated with the differences between these two instruments. The analyte concentration associated with each difference spectrum in $\mathbf{M}$ is zero, since $\mathbf{X}_1$ and $\mathbf{X}_2$ have the same analyte concentrations. In this case, minimizing $\frac{1}{2}||\mathbf{X}_\eta \mathbf{b} - \mathbf{y}_\eta||_2^2$ in Eq. 7 amounts to pointing $\mathbf{b}$ away from the space spanned by instrument noise.

Previous studies have shown that wavelength selection alone can perform calibration maintenance and transfer.[29–32] By using Eq. 7, we hope to immunize the primary calibration model against spectral noise while simultaneously performing wavelength selection. This augmentation forms the basis of many recent calibration transfer and maintenance methods[33–35] including augmented classical least-squares procedures, which decompose spectra into pure-component concentrations and pure-component spectra.[36–40] More recently, a comprehensive review of two-norm and one-norm penalties for sparse multivariate calibration and maintenance was undertaken by Kalivas.[23]

Sparse Principal Component Regression and Partial Least Squares. The dominant regression algorithms in spectroscopy are PCR and PLS.

PCR and PLS can both be interpreted as the decomposition of spectra into matrix factors, i.e., $\mathbf{X} = \mathbf{URV}^T$. The matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and have unit length (orthonormal). The columns $\mathbf{u}_1$ and $\mathbf{v}_1$ are often referred to as the score and loading vectors, respectively. For PCR, the matrix $\mathbf{R}$ is diagonal and consists of singular values. For PLS, $\mathbf{R}$ is bidiagonal.[41] PCR, as its acronym suggests, utilizes principal component analysis (PCA) to compute the matrix factors.

The matrix decomposition $\mathbf{X} = \mathbf{URV}^T$ allows one to easily compute the pseudo-inverse of $\mathbf{X}$, whereby $\mathbf{X}^+ = \mathbf{VR}^+\mathbf{U}^T$. The ordinary least-squares solution is written as the linear combination of all the loading vectors $\mathbf{b} = \mathbf{X}^+\mathbf{y} = \mathbf{V}\alpha = \mathbf{v}_1\alpha_1 + \ldots + \mathbf{v}_n\alpha_n$, where $\alpha = \mathbf{R}^+\mathbf{U}^T\mathbf{y}$. PCR and PLS create small-norm solutions by projecting $\mathbf{y}$ onto a lower-dimensional subspace spanned by the first k loading vectors. This results in a solution that uses only the first k loading vector, in which $\mathbf{b} = \mathbf{v}_1\alpha_1 + \ldots = \mathbf{v}_k\alpha_k$. Sparse versions of PCR and PLS project but also "sparsify" the loading vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$, meaning the resulting linear combination $\mathbf{b}$ is sparse as well. There are many approaches that create sparse loading vectors,[42–49] but most of them qualitatively share a similar mechanism for coefficient suppression in the score vectors: they append the one-norm and two-norm penalty terms of the elastic net in Eq. 5 to the PCR or PLS optimization machinery.

Meta-Feature Selection. The spectra associated with a set of samples are often measured in many modalities, e.g., different excitation sources, detectors, channels, instruments, etc. As a result, a larger heterogeneous set of spectra can be concatenated vertically from spectral blocks, i.e., $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$, where $\mathbf{X}_i$ is an $m \times n_1$ matrix of spectra measured under a particular set of modalities. The total number of wavelengths $n = n_1 + \ldots + n_K$ across all blocks can be quite large, and performing wavelength selection on all n wavelengths by using LASSO can be daunting. To reduce the computational burden, one might want to perform some type of data dimension-reduction technique (wavelets, PCA, PLS, etc.) on each spectral block. If PCA or PLS is used, for example, then the ith spectral block $\mathbf{X}_i$ can be decomposed ($\mathbf{X}_i = \mathbf{U}_i\mathbf{R}_i\mathbf{V}_i^T$), where the n_i columns of $\mathbf{X}_i$ are replaced with the first $r_i (r_i \ll n_i)$ score vectors of $\mathbf{U}_i$. Instead of performing wavelength selection on $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_K]$, one can perform feature selection on the collection of score vectors $\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_K]$.

Alternatively, instead of compressing the spectra within each spectral block, one might want to suppress an entire spectral block of wavelengths. The group LASSO can perform suppression at the group level.[50] If the group sizes are all one, then the group LASSO reduces to the original LASSO. As a result, one can use the group LASSO to do "spectral block" selection where one could identify, say, which excitation sources and channels in tandem, contribute most to the prediction of an analyte.

LASSO APPROACHES

Since the original LASSO paper, many algorithms have been developed to solve the one-norm penalty problem. Most algorithms have been minor variants of the original LASSO algorithm, but a few have offered real insights into how regression coefficients can be suppressed or "zeroed out". In this section, we highlight some of the more insightful LASSO approaches. First, we give a brief illustration of the simplest and one of the earliest LASSO approaches: the shooting algorithm of Fu.[6]

Shooting Algorithm. For ease of illustration, we examine Eq. 2, using only one wavelength ($n = 1$). In this case, the data matrix $\mathbf{X}$ gets simplified to a vector of scalar values $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$, and we minimize a function of only one variable:

$$\min_b \frac{1}{2}\sum_{i=1}^{m}(x_i b - y_i)^2 - \lambda|b| \quad (8)$$

Although Eq. 8 is non-differentiable at $b = 0$, we can still minimize the objective by setting the derivative equal to zero for $b \neq 0$:

$$\sum_{i=1}^{m}(x_i b - y_i)x_i + \lambda\ \mathrm{sign}(b) = 0$$
$$\Rightarrow ab - c = S(\lambda) \quad (9)$$

(The minimization of non-differentiable functions is often treated by using subgradients.[51]) The scalars $a = \mathbf{x}^T\mathbf{x}$ and $c = \mathbf{x}^T\mathbf{y}$ in Eq. 9 represent the positive slope and intercept, respectively, of the line, while $S(\lambda) = -\lambda\mathrm{sign}(b)$ represents a step function. In Eq. 9, the optimal value of b occurs where the line $ab - c$ intersects the step function $S(\lambda)$, and the intersection can occur in one of three ways (see Fig. 1). All we really

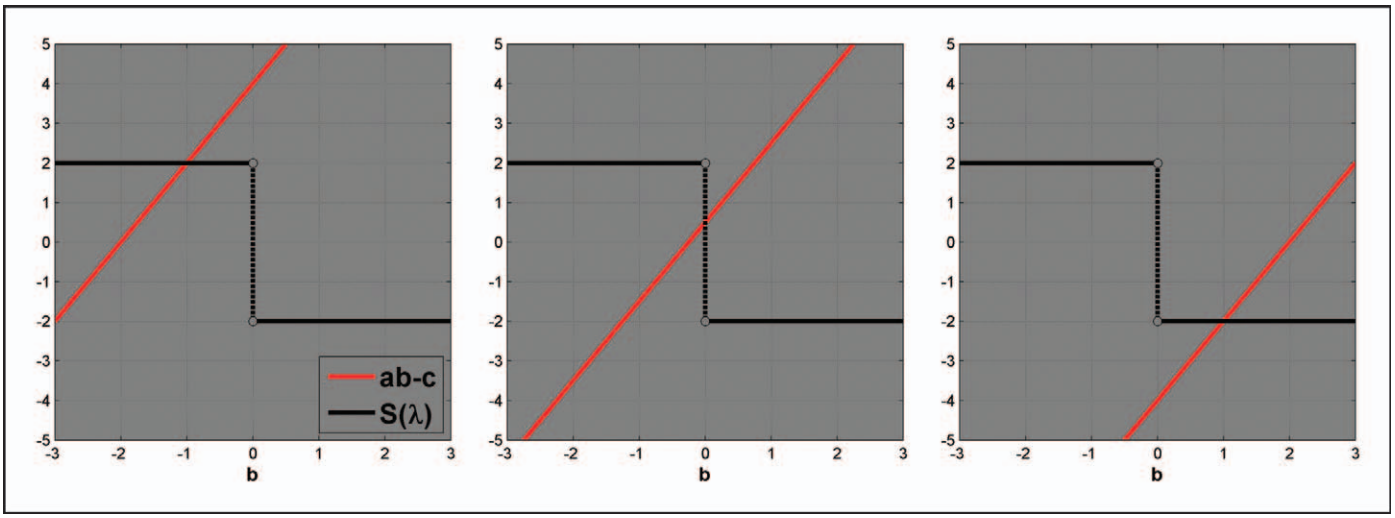

FIG. 1. For the shooting algorithm, the three ways that the line ab − c can intersect the step-like function S(λ) = λ sign(b). Sparsity, or suppression of b, to zero only occurs when the line intersects the vertical line segment in the middle subplot.

care about is the intercept c. If c is outside the interval $[-\lambda, \lambda]$, as shown in the left and right subplots of Fig. 1, we solve for b in Eq. 9, where $b = [S(\lambda) + c]/a$. However, in the middle subplot, the intersection between the line $ab - c$ and $S(\lambda)$ occurs within the interval $[-\lambda, \lambda]$ on the y-axis, at $b = 0$. The larger λ is, the longer the length of the interval on the vertical axis and the greater the likelihood that b will be "suppressed", or set to zero.

In the multivariate case, when the number of wavelengths is greater than one, the shooting algorithm starts with an initial solution, often a regression vector obtained by PLS or PCR. Next, a coordinate descent approach is applied: One cycles through each regression coefficient b_j in turn, minimizing it with the aforementioned intersection approach while keeping all of the other regression coefficients fixed, i.e., b_i, $i = 1, \ldots, n$, $i \neq j$.[8] In effect, the shooting algorithm "warm starts" with a non-sparse solution and iteratively adjusts the value of regression coefficients, setting many of them to zero in the process. Since the introduction of the shooting algorithm, an extremely efficient extension, colloquially referred to as "glmnet," was developed by Friedman et al.[8]

Least-Angle Regression. Suppose we solve the LASSO problem in Eq. 2 or Eq. 5 with a large value λ, such that all of the regression coefficients are set to zero, the sparsest calibration model of all. We then reduce the value of λ by a very small amount and solve the LASSO problem again. We repeat this process for a total of N times until λ is zero. This procedure generates a sequence of λ values $\lambda_0, \lambda_1, \ldots, \lambda_N$ and a set of corresponding regression vectors $\mathbf{b}^{[0]}$, $\mathbf{b}^{[1]}, \ldots, \mathbf{b}^{[N]}$, where $\mathbf{b}^{[1]}$ is the regression vector associated with its corresponding penalty parameter λ_1. However, if the change between the adjacent values of λ_1 and λ_{i+1} is small enough, then the number and position of the non-zero coefficients between the corresponding regression vectors $\mathbf{b}^{[i]}$ and $\mathbf{b}^{[i+1]}$ might not change. Hence, there will be a great deal of wasted computation for no qualitative change in solution.

Least-angle regression (LAR)[7] calculates the largest jump possible between adjacent values λ_1 and λ_{i+1}, such that $\mathbf{b}^{[i+1]}$ is the same as $\mathbf{b}^{[i]}$, except for one additional non-zero coefficient at another position. Hence, LAR "cold starts" with the most sparse regression vector $\mathbf{b}^{[0]}$ (a regression vector of all zeros) and builds, in a bottom-up fashion, n additional regression vectors $\mathbf{b}^{[1]}, \ldots, \mathbf{b}^{[n]}$ where $\mathbf{b}^{[1]}$ contains i non-zero coefficients. (The total number N of LAR iterations is $N = n$). The final LAR iterate $\mathbf{b}^{[n]}$ corresponds to the ordinary least-squares solution. There are two primary LAR variants, ordinary LAR or LAR with the LASSO modification. In the ordinary LAR approach, wavelength features are added in a greedy fashion—one feature at a time. However, the LASSO modification allows features to be added or removed. As a result, the total number N of LAR iterations will be much greater than n in the modified case. Unlike the shooting algorithm, LAR does not require λ as an input. As a result, this makes LAR easier to use than other LASSO methods. For example, for a given λ value in the shooting algorithm, one does not know a priori how many non-zero regression coefficients will result after calibration.

The primary computational overhead associated with LAR is that, at the ith iteration, an $i \times i$ linear system must be solved. Hence, for data sets with hundreds or thousands of wavelengths, the burden is minimal when i is small and maximal when i is large. Therefore, one might want to seek LASSO alternatives that are not so computationally burdensome. Iteratively reweighted least-squares schemes (IRLS) provide one such alternative.

Iteratively Reweighted Least Squares. Suppose we have a priori information regarding unequal weighting for each element in $\mathbf{b}$. For example, assume that the regression coefficients $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n$ in $\mathbf{b}$ are random variables that follow a normal distribution with

probability density function $\rho(\mathbf{b}) = \text{const} \times \exp(-\frac{1}{2}\mathbf{b}^T\mathbf{C}_b^{-1}\mathbf{b})$ and are independently and identically distributed with $\mathbf{C}_b$, the corresponding covariance matrix. The maximum likelihood estimate (MLE) from statistics is then found by minimizing:[52]

$$\min_{\mathbf{b}} \frac{1}{2}\left\|\mathbf{Xb}-\mathbf{y}\right\|_2^2 + \frac{1}{2}\mathbf{b}^T\mathbf{C}_b^{-1}\mathbf{b} \quad (10)$$

However, the MLE approach requires a priori estimates of $\mathbf{C}_b$, which are rarely known in practice. To make the estimation easier, we assume that the covariance matrix is a diagonal matrix of variances $\mathbf{C}_b = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, with σ_1 representing the spectral noise associated with the ith wavelength. In the absence of measuring the same spectra repeatedly and calculating the sample standard deviation, we can instead use the magnitude of the regression coefficient, derived from an initial least-squares estimate, as a proxy for σ_1.

Let $\mathbf{b}^{[0]} = (b_1^{[0]}, b_2^{[0]}, \ldots, b_n^{[0]})^T$ be an initial least-squares estimate from a standard MC procedure such as PLS, PCR, or TR. If the inverse of the diagonal covariance matrix is expressed as:

$$\mathbf{C}_b^{-1} = \lambda^2\mathbf{L}, \text{ where } \mathbf{L} = \text{diag}\left(\frac{1}{\left|b_1^{[0]}\right|}, \frac{1}{\left|b_2^{[0]}\right|}, \ldots, \frac{1}{\left|b_N^{[0]}\right|}\right) \quad (11)$$

then the MLE problem in Eq. 10 can be expressed as:

$$\min_{\mathbf{b}^{[1]}} \frac{1}{2}\left\|\mathbf{Xb}^{[1]}-\mathbf{y}\right\|_2^2 + \frac{1}{2}\lambda^2\left\|\mathbf{Lb}^{[1]}\right\|_2^2 \quad (12)$$

The solution update $\mathbf{b}^{[1]} = (b_1^{[1]}, b_2^{[1]}, \ldots, b_n^{[1]})^T$ in Eq. 12 is a smaller-norm version of the initial estimate $\mathbf{b}^{[0]}$, since the coefficient update $b_i^{[1]}$ is penalized by $\mathbf{L}_{ii} = 1/|b_i^{[k]}|$ in Eq. 11. If the magnitude of the previous update $|b_i^{[0]}|$ is small (has little variance), then $\mathbf{L}_{ii}$ will be large, and $b_i^{[1]}$ will be driven to zero. Such an approach was successfully used in wavelength selection by Ottaway et al.[53]

The reweighting process of Eq. 12 can be repeated iteratively to obtain a sequence $\mathbf{b}^{[0]}, \mathbf{b}^{[1]}, \mathbf{b}^{[2]}, \ldots$ of regression vectors. At the kth iteration, a regression coefficient $b_i^{[k]}$ will be set to zero if its magnitude is below some threshold, e.g., $|b_i^{[k]}| < \tau$, where $\tau = 10^{-8}$. The iterative nature of Eq. 12 can be recast within a larger framework of IRLS schemes[54–57] that generate sparse regression vectors for wavelength selection—see Table I. Here, the regression coefficients from a classical MC method are used (as described next) to create a diagonal matrix $\mathbf{F}^{[k]} = \text{diag}(f_1^{[k]}, f_2^{[k]}, \ldots, f_n^{[k]})$, which rescales each column of $\mathbf{X}$.

TABLE I. Iteratively reweighted least-squares scheme for wavelength selection.

Step	Instruction(s)
0	Solve $\mathbf{Xb}^{[0]} = \mathbf{y}$ for $\mathbf{b}^{[0]}$, using PLS, PCR, or TR; set $k = 1$
1	Form scaling matrix $\mathbf{F}^{[k]} = \text{diag}(f_1^{[k]}, f_2^{[k]}, \ldots, f_n^{[k]})$
2	Solve $\mathbf{\Phi}^{[k]}\boldsymbol{\beta}^{[k]} = \mathbf{y}$ by using PLS, PCR, or TR for $\boldsymbol{\beta}^{[k]}$, where $\mathbf{\Phi}^{[k]} = \mathbf{XF}^{[k]}$
3	Recover $\mathbf{b}^{[k]}$ by using back-substitution $\mathbf{b}^{[k]} = \mathbf{F}^{[k]}\boldsymbol{\beta}^{[k]}$
4	Set $k = k + 1$ and go to Step 1.

Each element $f_i^{[k]}$ of the diagonal scaling matrix is a mathematical expression involving the regression coefficient from the previous iteration, i.e., $b_i^{[k-1]}$. If the diagonal element $f_i^{[k]}$ is approximately zero, then the ith wavelength of the spectra can effectively be ignored. If k diagonal entries of $\mathbf{F}^{[k]}$ are zero, then $\mathbf{\Phi}^{[k]} = \mathbf{XF}^{[k]}$ contains k columns that are all approximately zero, and they can be removed from consideration, where $\mathbf{\Phi}^{[k]}$ is of dimension $m \times (n - k)$. The diagonal scaling element at the kth iteration is defined as:

$$f_i^{[k]} = \left|b_i^{[k-1]}\right| \text{ or } \mathbf{F}^{[k]} = \text{diag}\left(\left|b_1^{[k-1]}\right|, \left|b_2^{[k-1]}\right|, \ldots, \left|b_n^{[k-1]}\right|\right) \quad (13)$$

Equation 13 is a special case of an iterative technique, developed in the signal processing community, called the focal underdetermined system solution algorithm, or FOCUSS.[55,56]

In the signal- and image-processing communities, the sparse methodologies outlined in this paper are often referred to as "compressive sensing", and IRLS algorithms of the type outlined in Table I are commonplace.[57–59] In the chemometrics literature, diagonal weighting schemes using a variety of mathematical expression for $f_i^{[k]}$ have been successfully employed for wavelength selection.[54,60,61]

The sparse iterative framework in Table I is appealing in that no sophisticated optimization-based solvers are required; sparse calibration models can be obtained by the simple recycling of regression coefficients obtained from conventional MC methods. Moreover, the sparse iterative framework allows for even faster algorithms by taking advantage of advances within conventional MC methods. For example, randomized algorithms (e.g., randomized PCA) construct approximate matrix factorizations of the data by using random sampling to quickly capture the subspace that explains the dominant variability, or "action", of a matrix.[62–65] As a result, these randomized algorithms can handle massive data sets, unlike conventional algorithms (such as ordinary or deterministic PCA). Extensive numerical experiments have shown that these algorithms often outperform their deterministic counterparts in terms of accuracy, speed, and robustness. Hence, in Table I, one can use PCR for the regression engine in step 2, with PCA being replaced with a randomized PCA.

SPARSE METHODS FOR SAMPLE SELECTION

In this section, we define sparse methods for sample selection, i.e., finding a predictive subset of samples. Instead of generating an n-dimensional regression vector whereby each regression coefficient has a one-to-one correspondence with a particular wavelength, we generate an m-dimensional regression vector, whereby each regression coefficient has a one-to-one correspondence with a particular sample or spectrum. The sample selection approach we propose involves the recast-

ing of support-vector regression (SVR) as a LASSO problem.

SVR, like other regression approaches, strives to find an optimal hyperplane of fit by minimizing the residuals $r_i = \mathbf{x}_i^T\mathbf{b} - y_i$, but unlike other regression approaches, the optimization machinery used to define the hyperplane is qualitatively different.[66–68] There are many SVR variants, and the one we opt for allows for a direct link with the elastic net variant of the LASSO.

The linear SVR variant we use solves the following unconstrained minimization problem:

$$\min_{\mathbf{a}} \tfrac{1}{2}\mathbf{a}^T(\mathbf{K} + \eta^2\mathbf{I}_m)\mathbf{a} - \mathbf{a}^T\mathbf{y} + \lambda\|\mathbf{a}\|_1 \tag{14}$$

where $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ is called the kernel matrix.[54,66–68] Instead of solving for the primal regression vector $\mathbf{b} = [b_1, b_2, \ldots, b_n]^T$ (one coefficient for each wavelength) as we did in the LASSO, we solve for the dual-regression vector $\mathbf{a} = [a_1, a_2, \ldots, a_n]^T$ (one coefficient for each sample). The primal regression vector is related to the dual solution $\mathbf{a}$ via a linear combination of the spectra:

$$\mathbf{b} = \mathbf{X}^T\mathbf{a} = \sum_{i=1}^{m} a_i\mathbf{x}_i = a_1\mathbf{x}_1 + \cdots + a_n\mathbf{x}_n \tag{15}$$

Like the LASSO, the one-norm penalty in Eq. 14 creates a sparse dual-regression vector $\mathbf{a}$. The samples associated with the non-zero coefficients a_i are referred to as the "support" vectors, since they alone contribute to the summation in Eq. 15. Note that the dual-regression vector is sparse and not the primal regression vector, since it is a linear combination of dense spectral measurements. In addition, the SVR formulation in Eq. 14 can easily be generalized to nonlinear regression (via a nonlinear kernel[66,67]), but we restrict ourselves in this paper to the linear-regression setting.

To see the connection between SVR and LASSO, we note that the kernel matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ is symmetric, positive semi-definite (all of its eigenvalues are nonnegative), and this property allows us to take fractional powers of $\mathbf{K}$ via singular-value decomposition (SVD): $\mathbf{K}^p = \mathbf{U}\mathbf{\Sigma}^p\mathbf{U}^T$, where $\mathbf{U}$ is a matrix of singular vectors (the orthogonal score vectors of $\mathbf{X}$), and $\mathbf{\Sigma}$ is a diagonal matrix of singular values. Using the variable transformation of Franklin:[69]

$$\bar{\mathbf{K}} = \mathbf{K}^{\frac{1}{2}} \text{ and } \bar{\mathbf{y}} = \mathbf{K}^{-\frac{1}{2}}\mathbf{y} \tag{16}$$

we can rewrite Eq. 14 as:[54]

$$\min_{\mathbf{a}} \tfrac{1}{2}\left\|\mathbf{K}_\eta\mathbf{a} - \mathbf{y}_\eta\right\|_2^2 + \lambda\|\mathbf{a}\|_1, \text{ where } \mathbf{K}_\eta = \begin{bmatrix}\bar{\mathbf{K}} \\ \eta\mathbf{I}_m\end{bmatrix} \text{ and } \mathbf{y}_\eta = \begin{bmatrix}\bar{\mathbf{y}} \\ 0_m\end{bmatrix} \tag{17}$$

As a result, Eq. 17 is the same as the elastic-net variant of the LASSO in Eq. 6, except for the substitutions: $\bar{\mathbf{K}}_\eta$ for $\mathbf{X}$, $\bar{\mathbf{y}}_\eta$ for $\mathbf{y}$ and $\mathbf{a}$ for $\mathbf{b}$. We denote Eq. 17 as SVR–LASSO. When λ is zero, all of the samples are support vectors and SVR simplifies to what is known as kernel-ridge regression.[66] Kernel-ridge regression amounts to the solution of the linear system $\bar{\mathbf{K}}\mathbf{a} = \bar{\mathbf{y}}$ via TR, where the primal regression vector is recovered via the relation $\mathbf{b} = \mathbf{X}^T\mathbf{a}$ in Eq. 15. (Note that the solution $\mathbf{a}$ to either linear systems $\mathbf{K}\mathbf{a} = \mathbf{y}$ or $\bar{\mathbf{K}}\mathbf{a} = \bar{\mathbf{b}}$ is the same.)

The functional equivalence between SVR–LASSO in Eq. 17 and the LASSO variant in Eq. 6 means that all of the LASSO algorithms used for wavelength selection can be re-appropriated for sample selection. For example, LAR can now be applied to Eq. 17 to generate a sequence of support-vector solutions $\mathbf{a}^{[0]}, \mathbf{a}^{[1]}, \ldots, \mathbf{a}^{[m]}$, where $\mathbf{a}^{[i]}$ contains i non-zero coefficients. Alternatively, one can use the iterative framework of IRLS to solve Eq. 17 (see Table II) for sample selection purposes.

REGRESSION EXAMPLES AND PRACTICAL IMPLEMENATION

In this section, we present examples in which classical MC methods are compared with their sparse counterparts. We also compare sparse methods for wavelength selection and sample selection.

We show the profile of regression coefficients as a function of wavelength and the RMSE values as a function of the number of wavelengths or the number of latent vectors. To demonstrate the simplicity of the regression methods for wavelength and sample selection, code will be available at www.hpc.unm.edu/~andriese. All of the software was written in MATLAB, release 2010b.

Data Sets. We here examine three data sets: corn,[70] wheat,[71] and blood[72] for purposes of wavelength and sample selection.

The corn data set consists of 80 samples of corn, with 700 absorbances measured from 1000 to 2498 nm, at 2 nm intervals on three near-infrared (NIR) spectrometers, designated m5, mp5, and mp6. Reference values are provided for oil, protein, starch, and moisture content. Protein content is the prediction property studied in this paper, and the spectra measured on instrument m5 serves as the primary calibration set. Unlike the other two data sets, there are no designated calibration and validation sets. To construct such sets, we arbitrarily split the 80 samples into halves, with the calibration data consisting of the first 40 samples, with the remaining samples composing the validation data.

The wheat data set consists of 884 spectral samples (777 calibration and 107 validation samples) of whole-grain Canadian wheat, measured by diffuse reflectance spectroscopy. The calibration samples represent samples grown in years 1998 and 2000–2005. The validation samples were grown in 1999 and are quite separate from the calibration samples. There are 1038 wavelengths from 400 to 2499 nm, at 2 nm intervals. There are many references associated with this data set, but we are only interested in percentage protein content for each sample. This data set was featured in the "NIR Shootout" data of the 2008 International Diffuse Reflectance Conference in Chalmersburg, PA.

The blood data set consists of 553 blood samples (472 calibration and 81 validation samples), obtained from 13 healthy human volunteers. Reference values for blood glucose range between 30 and 500 mg/dL. Spectroscopic data were collected with a Fourier transform NIR spectrometer, with 260 absorbances measured between 4000 and 8000 cm^{-1}. The blood samples were pumped through a borosilicate flow cell with a nominal pathlength of 1 mm, which was

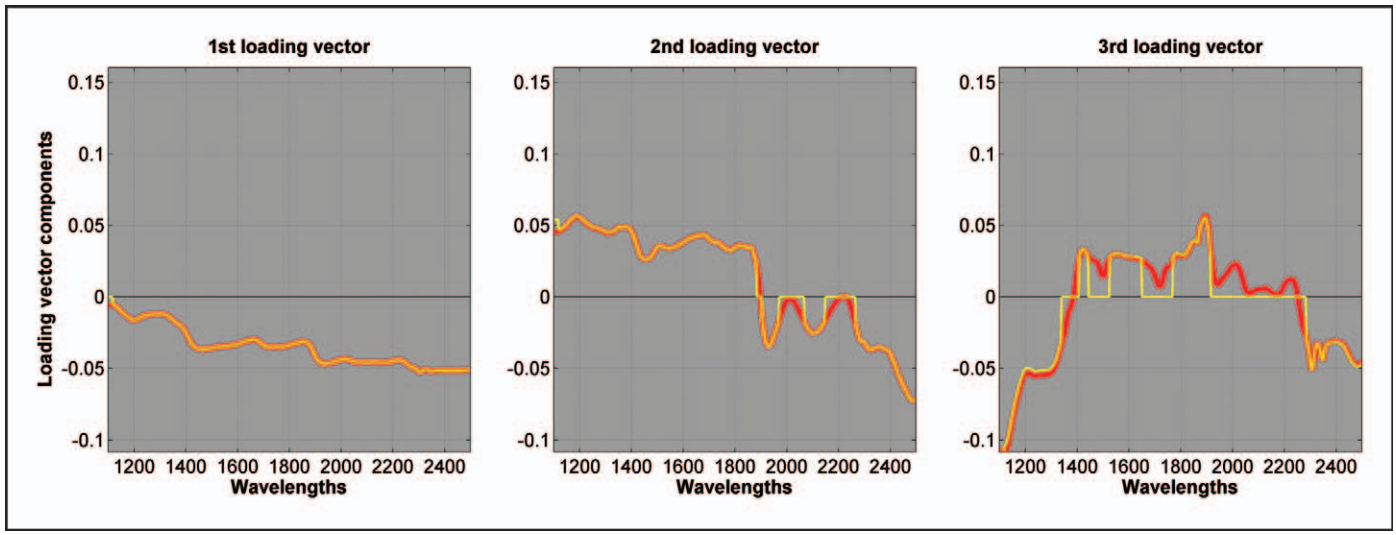

FIG. 2. Comparison of the first three loading vectors, computed by ordinary PCA (red), and the first three loading vectors, computed by sparse PCA (yellow).

temperature controlled at 34 ± 0.5 °C. Only spectra in the regions 4200–4950 and 5400–7200 cm^{-1} were analyzed (the truncated region 4950–5400 cm^{-1} is associated with the water band), for a total of 164 wavenumbers.

For all three data sets, no modifications or preprocessing treatments (other than the initial mean centering) were made. In the statistics literature, it is commonplace to scale the data to have unit variance across variables, especially when the variables are measured in different units. However, in spectroscopic applications where all of the variables are measured on the same scale (e.g., absorbance units), there is no need to standardize the variables. The data sets analyzed here were not scaled to have unit variance across wavelengths.

Sparse Principal Component Analysis. For an illustration of sparse PCA, we compute the first three loading vectors associated with the calibration samples of the corn data set. The sparse PCA implementation that we used is based on a gradient-based optimization scheme called G*Power.[46] (G*Power is an acronym for gradient power analyses.) In Fig. 2, we compare the first three loading vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, computed by ordinary PCA (red curves), with the first three sparse loading vectors $\tilde{\mathbf{v}}_1$, $\tilde{\mathbf{v}}_2$, $\tilde{\mathbf{v}}_3$, computed by G*Power (yellow curves). In G*Power, one can increase the sparsity per loading vector by changing a tuning parameter called the sparsity weighting factor.[46] (Here, the sparsity weighting factor was set to 0.11 for each loading vector.) The first sparse PCA loading vector is practically identical to the loading vector associated with ordinary PCA. However, the second and third G*Power loading vectors show increasing amounts of sparsity. If one were to use only these three loading vectors $\tilde{\mathbf{v}}_1$, $\tilde{\mathbf{v}}_2$, $\tilde{\mathbf{v}}_3$ for PCR, then the regression vector (a linear combination of these vectors) would still not be sparse, since the zero-valued elements across all three loading vectors do not overlap. For truly-sparse loading vectors, one would have to increase the sparsity weighting factors.

Shooting Algorithm. The shooting algorithm solves the elastic net variant of the LASSO in Eq. 5. This algorithm requires a "guess" for the initial regression vector, and we choose the TR solution of Eq. 1 with the two-norm penalty parameter fixed at $\eta = 10^{-6}$. Here, we also use the calibration samples of the corn data set. Keeping $\eta = 10^{-6}$ fixed, we then choose three one-norm penalty parameters: $\lambda_1 = 5.9 \times 10^{-6}$, $\lambda_2 = 5.9 \times 10^{-5}$, $\lambda_3 = 5.9 \times 10^{-4}$. (These λ values correspond to λ_{max} being divided by 100 000, 10 000, and 1000, respectively, where $\lambda_{max} = 5.9$ is the effective upper boundary on λ, above which all regression coefficients are set to zero.) We then compare the TR model vector $\mathbf{b}_{TR}$, with the three corresponding LASSO model vectors $\mathbf{b}^{[1]}$, $\mathbf{b}^{[2]}$, $\mathbf{b}^{[3]}$, associated with λ_1, λ_2, λ_3. The results are shown in Fig. 3.

In Fig. 3A, all four regression vectors are shown across all wavelengths. As the one-norm penalty parameters increase, the regression vectors become sparser. Figure 3B is the same as Fig. 3A, except that the *x*-axis has been restricted between 1850 and 2200 nm. For the largest one-norm penalty, the non-zero coefficient profiles coalesce around two wavelength intervals or bands.

The shooting algorithm is also representative of any algorithm that requires λ as an input: For a given λ value, one does not know a priori the number of non-zero regression coefficients that will result after calibration. A two-fold change in λ could result in a drastic change in the number of wavelengths. Hence, it would be more spectroscopically intuitive to vary the number of wavelengths as opposed to varying λ.

Wavelength Selection via Least-Angle Regression and Iterative Reweighting Least Squares. We compare two MC wavelength-selection methods, IRLS by using PLS in Table I and LAR against PLS. In the case of LAR, we use the DTU-LAR, a LAR implementation from the SpaSM (sparse statistical modeling) toolbox, developed at DTU (the Danish acronym for the Technical University of Denmark).[73] Ordinary

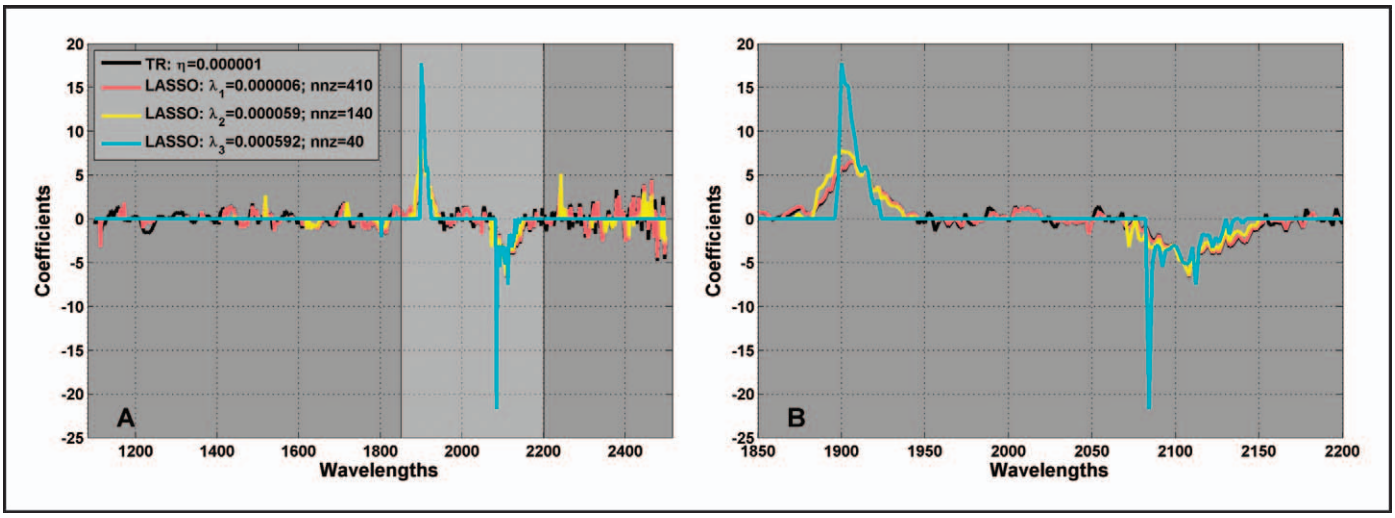

FIG. 3. The left subplot shows the value of the regression coefficients as a function of wavelength for the corn data set. In TR, all 700 regression coefficients are non-zero. For the LASSO with the shooting algorithm, the number of non-zero coefficients decreases as the one-norm penalty parameter increases. The right subplot is the same as the left except that the interval for the vertical x-axis has been restricted to [–1850, 2200].

LAR (as opposed to the LASSO modification of LAR) is used. The second wavelength selection is the IRLS framework of Table I, with PLS as the base MC method. We denote this method as IRLS–PLS. Five IRLS iterations are used.

In Fig. 4, the regression vectors for PLS, LAR, and IRLS–PLS, determined from the corn calibration set are shown. For PLS, we compute 40 regression vectors, such that the *i*th regression vector uses the first *i* latent vectors (Fig. 4A). By using LAR, 700 regression vectors are generated, such that the *i*th regression vector contains *i* non-zero regression coefficients (Fig. 4B). The regression vectors associated with small numbers of non-zero regression coefficients (or number of wavelengths used) are illustrated in white and have large amplitudes. As the sparsity decreases to the point where most coefficients are non-zero, the coefficient amplitudes are commensurate with that of PLS when using many latent vectors (the yellow curves in Fig. 4B). For IRLS–PLS, 40 regression vectors are shown (one for each number of latent vectors used) (Fig. 4C). Compared with ordinary PLS in Fig. 4A, the regression vectors for IRLS–PLS after five iterations in Fig. 4C are very sparse.

In Fig. 5, the RMSE results associated with the model vectors in Fig. 4 are shown. Figure 5A shows the RMSE values for LAR as a function of the number of wavelengths used in the calibration model. Here, both the minimum of both the RMSECV and RMSEV (RMSE of validation) occurs when 25 (out of 700) wavelengths are

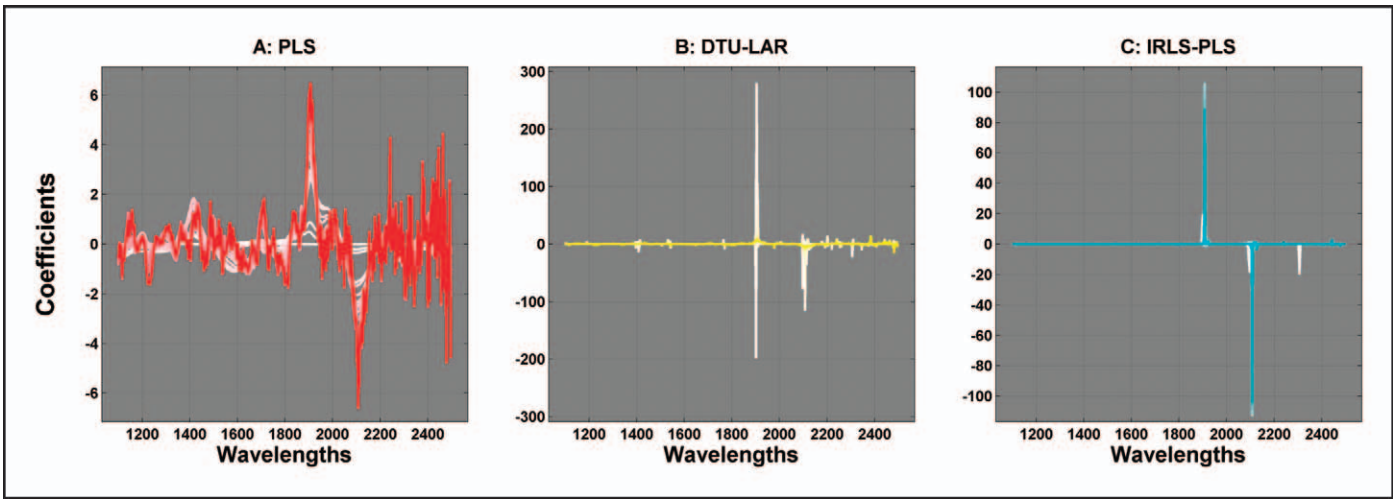

FIG. 4. The plot of regression coefficients as a function of wavelength for the corn data set across three methods: PLS, LAR, and IRLS–PLS.

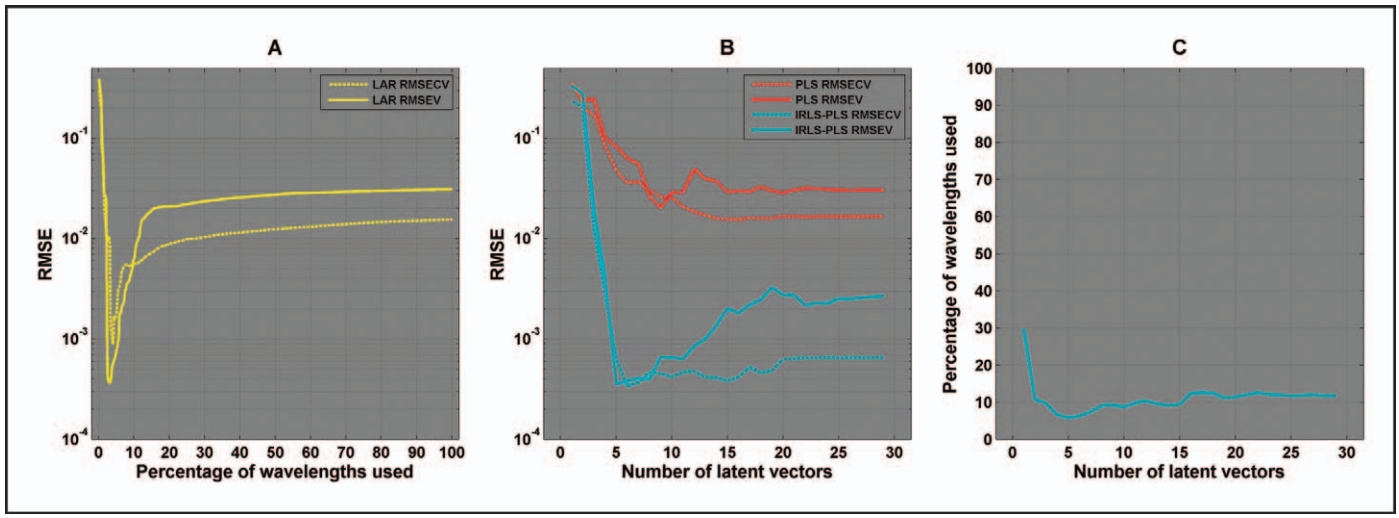

FIG. 5. Wavelength selection for the corn data set. (A) RMSECV and RMSEV values associated with least-angle regression (LAR). (B) RMSECV and RMSEV values associated with PLS (red) and IRLS–PLS (cyan). (C) For each number of latent vectors, the percentage of wavelengths used in the calibration model for IRLS–PLS.

used in the calibration model, or about 3.6% of the total number of wavelengths. Figure 5B displays the RMSE values for both ordinary PLS and IRLS–PLS after five iterations. In this case, the iterative improvement over PLS is significant. Figure 5C shows, for each number of latent vectors used in the calibration model for IRLS–PLS, the percentage of wavelengths used. For the corn data set, IRLS–PLS achieves a lower RMSE over PLS, using considerably fewer wavelengths.

Figure 6 displays the RMSE results for the wheat data set. The same description convention described in Fig. 5 is used here as well. Wavelength selection via LAR (Fig. 6A) and IRLS–PLS (Fig. 6B) confers no advantage in RMSE performance over ordinary PLS (Fig. 6B). However, a well-performing calibration model that is non-inferior to PLS requires only a small percentage of wavelengths, about 8% for LAR (Fig. 6A) and about 20% for IRLS–PLS (Fig. 6C). Evident in Figs. 6B and 6C is the trade-off between two types of parsimony, parsimony in the number of latent vectors and parsimony in the percentage of wavelengths used. Although IRLS–PLS performs as well as PLS when using fewer wavelengths, IRLS–PLS has to increase the number of latent vectors to achieve the same level of performance. The large discrepancy between RMSEV and RMSECV is due to the make-up of the calibration and validation samples: the calibration samples are from years 1998 and 2000–2005, while the validation samples are from the year 1999. For the blood data set in Fig. 7, the same description convention is used as in Figs. 5 and 6. As in the wheat data set, wavelength selection is non-inferior RMSE-wise to

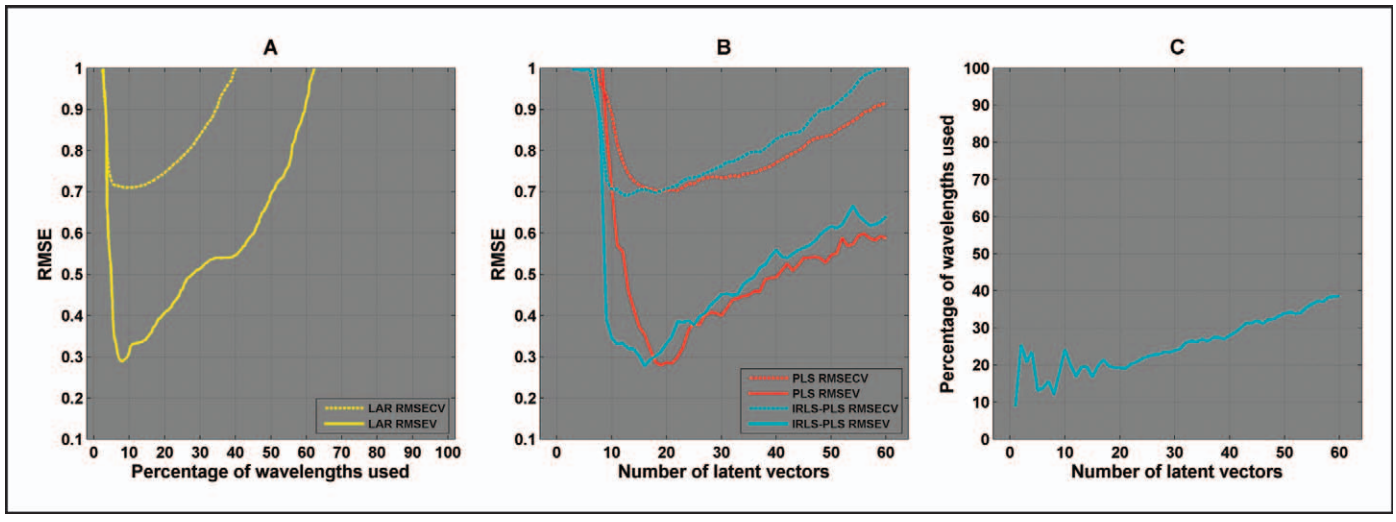

FIG. 6. Wavelength selection for the wheat data set. The same description convention is used as described in Fig. 5.

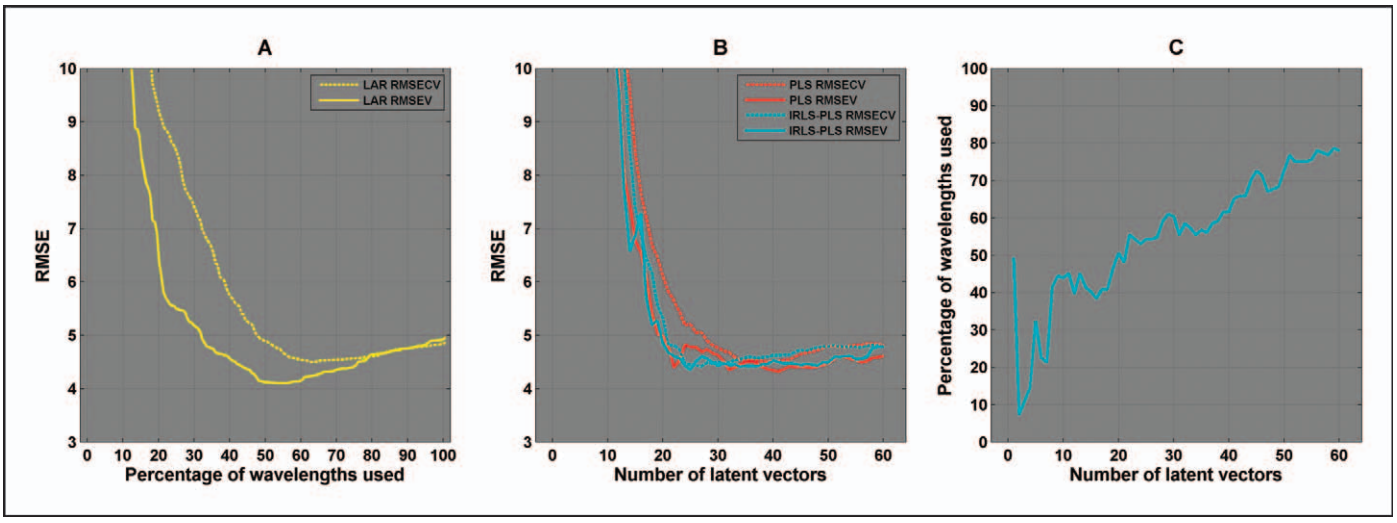

FIG. 7. Wavelength selection for the blood data set. The same description convention is used as described in Fig. 5.

PLS. Unlike both the wheat data and corn data sets, a larger percentage of wavelengths (about 50%) are needed by both LAR and IRLS–PLS to achieve commensurate performance with PLS. This increase in the percentage of wavelengths used is likely due to the typical heterogeneity of blood samples: 472 samples across 13 human volunteer samples containing significantly varying levels of hematocrit (the volume percentage of red blood cells in blood; the other blood components are plasma, white blood cells, and platelets).

For blood glucose, a particular type of scatter plot, called the Clarke error grid analysis[74] (CEGA) plot, is of diagnostic interest to clinicians. The CEGA plot divides the clinical accuracy of blood glucose (estimate, reference) coordinates into five regions: A, B, C, D, and E (see Figs. 10 and 11). Region A corresponds to sufficiently accurate estimates that are within 20% of their reference values. The coordinates in region B contain estimates that would not cause one to embark on inappropriate diabetes treatment. Region C contains estimates that would lead to inappropriate diabetes treatment. The coordinates in region D represent a failure to detect hyperglycemia (high blood sugar) or hypoglycemia (low blood sugar). Region E contains coordinates whose estimates would confuse hyperglycemia for hypoglycemia, and vice versa. Figure 10 contains the CEGA plots for PLS (using 25 latent vectors), LAR (using 30% of the wavelengths in the calibration model), and IRLS–PLS (using 25 latent vectors). (The parameters 25 and 30% were chosen arbitrarily; typically, these parameters would be chosen on the basis of some model selection criteria.) The CEGA plots report the (i) percentage of

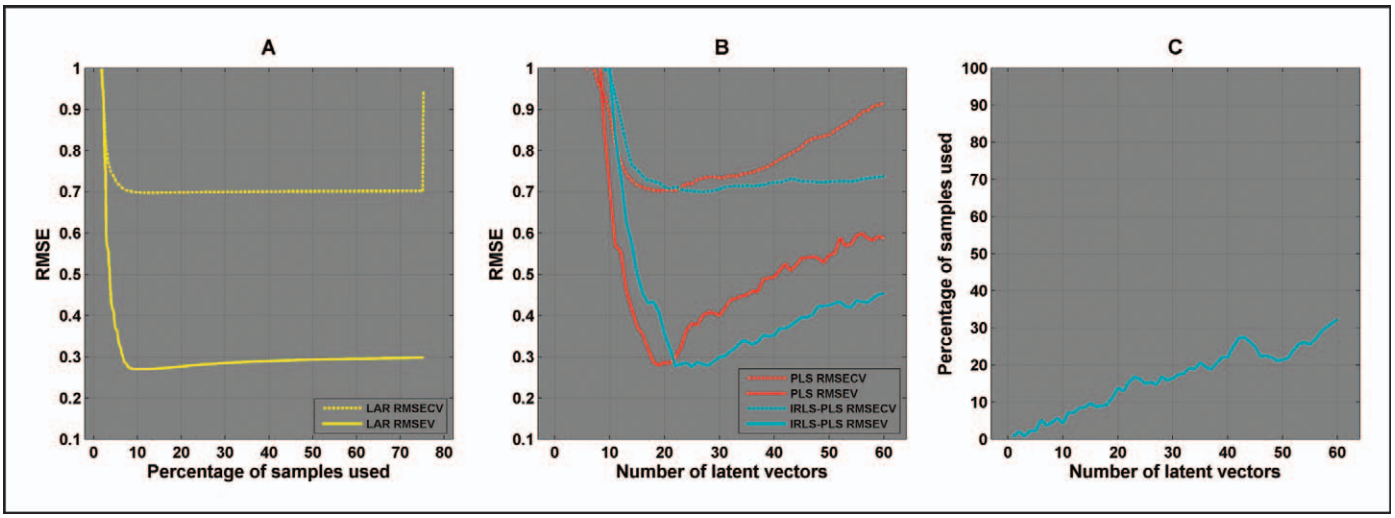

FIG. 8. Sample selection for the wheat data set. (A) RMSECV and RMSEV values associated with LAR. (B) RMSECV and RMSEV values associated with PLS (red) and IRLS–PLS (cyan). (C) For each number of latent vectors, the percentage of samples used in the calibration model for IRLS–PLS.

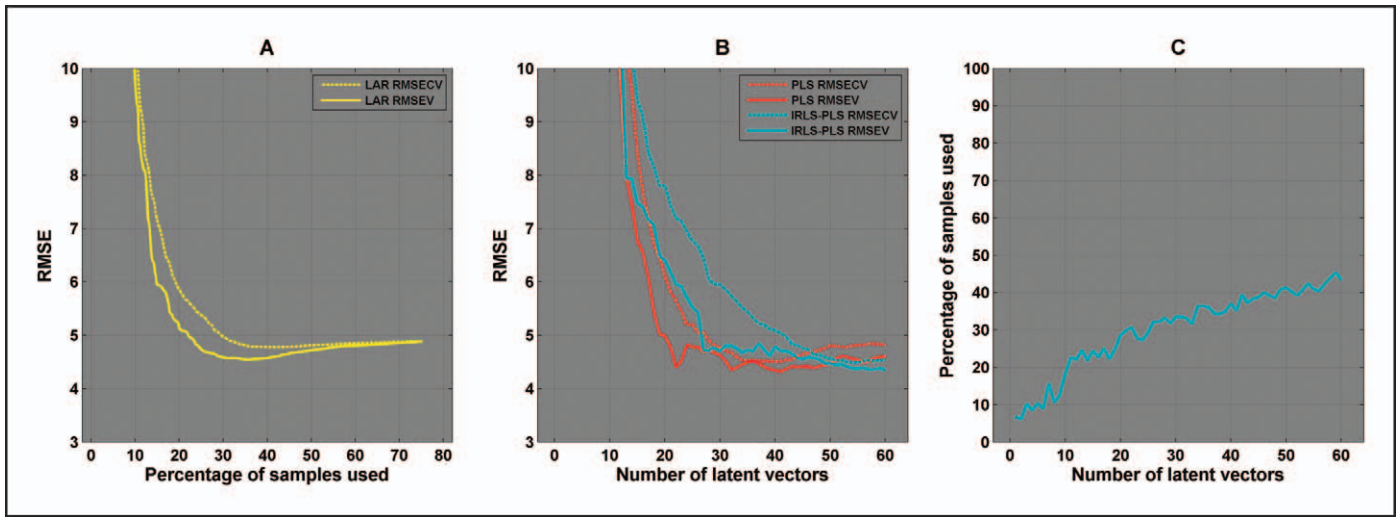

FIG. 9. Sample selection for the blood data set. The same description convention is used as described in Fig. 8.

coordinates in each region; (ii) the RMSE of validation; (iii) the slope, intercept and *R*-value of the line of fit; (iv) the number of validation samples; and (v) the percentage of wavelengths used in the calibration.

Sample Selection via Least-Angle Regression and Iterative Reweighting Least Squares. As was the case with wavelength selection, we compare LAR and IRLS–PLS against PLS. For IRLS–PLS, we follow the scheme in Table II, with PLS as the base MC method. We only examine the blood and wheat data set, since they have sufficiently large sample sizes. Figures 8 and 9 examine the RMSE performance of these three methods with respect to sample selection.

Figure 8 displays the RMSE results for the wheat data set. In Fig. 8A, the RMSE associated with LAR is shown as a function of the percentage of samples used in the calibration. The *x*-axis limit only goes up to 75%, since fourfold cross-validation was used; three-quarters of the data were used in each fold. The striking difference for LAR between sample and wavelength selection is the relative insensitivity of RMSE across a large swath of sample percentages used. In 10–70% of the samples used, the RMSE values are approximately the same. Figure 8B shows the RMSE performance for PLS and IRLS–PLS, with IRLS–PLS having commensurately the same performance as PLS. As was the case with wavelength selection, IRLS–PLS uses considerably fewer variables (samples in this case) than PLS, but commensurate performance is obtained only when more latent vectors are used. Figure 9 displays the RMSE performance for the blood data set. Here, LAR achieves the same level of RMSE performance as PLS. However, IRLS–PLS requires almost double the number of latent vectors to achieve the same level of RMSE performance as PLS, albeit with a fraction of the number of samples. Figure 11 contains the CEGA plots for PLS, LAR, and IRLS–PLS. It has the same description convention as

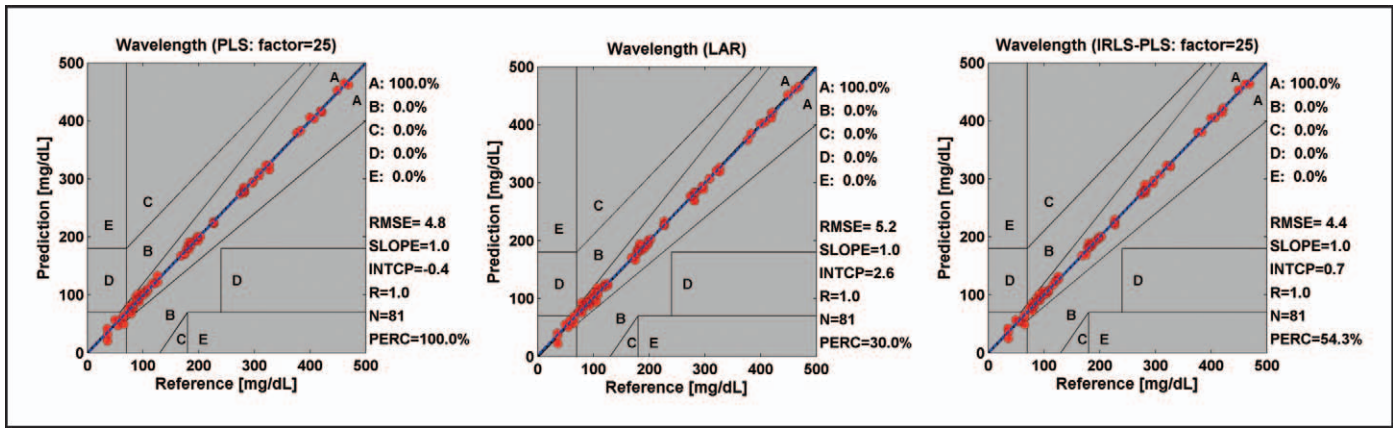

FIG. 10. Clarke error grid analysis (CEGA) plots associated with wavelength-selection methods: PLS (no wavelength selection) at 25 latent vectors (or factors), LAR at 30% of the wavelengths used in the calibration model, and IRLS–PLS at 25 factors.

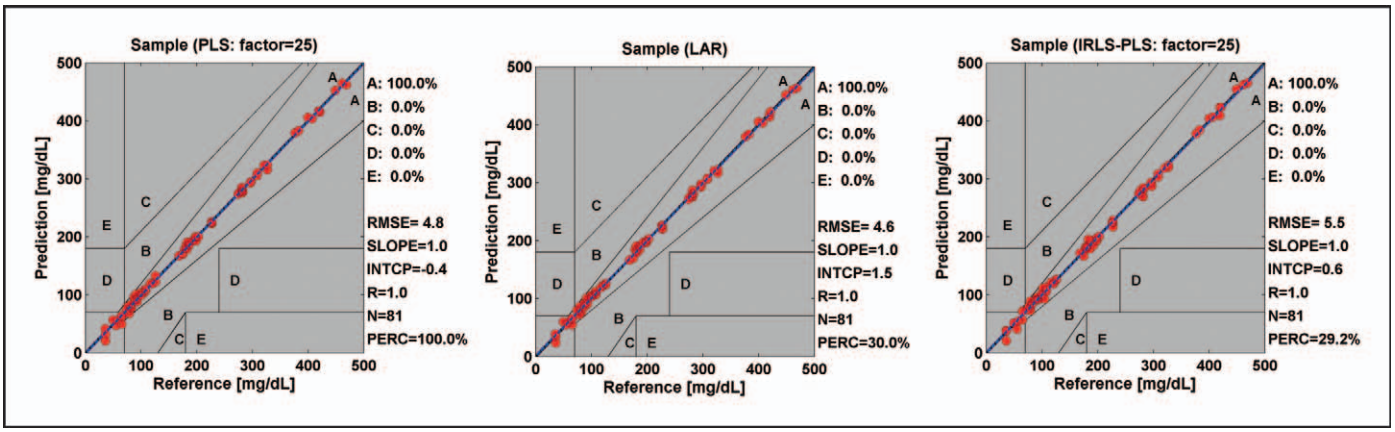

FIG. 11. Clarke error grid analysis (CEGA) plots associated with sample selection methods: PLS (no sample selection) at 25 latent vectors (or factors), LAR at 30% of the samples used in the calibration model, and IRLS–PLS at 25 factors.

Fig. 10, except that it reports the percentage of samples used in the calibration instead of the percentage of wavelengths.

CONCLUSION AND FUTURE WORK

We here reviewed some of the basic mechanisms and implementations of sparse methods, both for wavelength and sample selection. By using SVR, we adopted the perspective that wavelength and sample selection are actually two sides of the same coin. Furthermore, using IRLS-based methods, we show that sparse methods need not be overly complex to implement. For example, IRLS–PLS reuses an existing, off-the-shelf, PLS implementation in a simple iterative scheme to generate sparse models.

Compared with classical MC methods such as PLS, sparse methods are non-inferior (and, sometimes, considerably superior) in terms of RMSE performance. Moreover, only a fraction of the wavelengths or samples are needed for an adequate calibration model. We did not explore the alternating use of wavelength and sample selection procedures to construct a minimal spectral subset of samples and wavelengths.

Extensions to Classification. In the classification setting, say binary classification, the response variables are discrete, e.g., $\mathbf{y} = [y_1, y_2, \ldots, y_m]^T$, where the ith sample belongs to either the positive or negative class, or $y_i = \{-1, +1\}$. The regression algorithms discussed here can easily be repurposed for classification. For example, the prediction on a novel spectrum can be turned into a positive or negative class label via the signum function:

$$\text{sign}[f(\mathbf{z})], \text{ where } f(\mathbf{z}) = (\mathbf{z} - \bar{\mathrm{x}})^T\mathbf{b} + \bar{\mathrm{y}} \quad (18)$$

For classification, one can also employ support-vector machines (SVMs). The SVM does for classification what SVR does for regression. Like SVR, the prediction for a novel spectrum in a SVM involves only support vectors that correspond to non-zero coefficients of the vector $\mathbf{a} = [a_1, a_2, \ldots, a_m]^T$. However, in the classical SVM, the coefficients must be non-negative. On the other hand, least-squares generalizations of SVMs have been developed that loosen this non-negativity constraint and, as a result, the modified SVMs behave like SVR.[75]

Although SVMs are powerful classifiers, linear discriminant analysis (LDA) is still the dominant classification algorithm in chemometrics and spectroscopy. LDA projects spectra onto the most discriminative, low-dimensional subspace, and the classification is done in this reduced space. Recently, a sparse version of LDA was developed, such that one obtains not only a sparse regression vector, but also a set of sparse basis vectors that span the discriminative subspace.[76]

Massive Data Sets. Currently, most spectroscopic data sets are not yet sufficiently large. In other words, we are still applying sparse methods to data sets by using a standalone computing entity such as a desktop or laptop. However, spectral-imaging capabilities that currently measure thousands of spectral samples in a few minutes will soon give way to hundreds of thousands or millions of samples. As the spectral data sets get inexorably more massive, algorithms generating sparse regression vectors might only be feasible on distributed computing environments. In addition, for process monitoring, in which large quantities of spectra are continuously collected and analyzed over time, one might only have access

TABLE II. Iteratively reweighted least-squares scheme for sample selection.

Step	Instruction(s)
0	Solve $\mathbf{K}\mathbf{a}^{[0]} = \mathbf{y}$ for $\mathbf{a}^{[0]}$ by using PLS, PCR, or TR; set $k = 1$
1	Form scaling matrix $\mathbf{F}^{[k]} = \text{diag}(f_1^{[k]}, f_2^{[k]}, \ldots, f_n^{[k]})$
2	Solve $\mathbf{\Phi}^{[k]}\boldsymbol{\alpha}^{[k]} = \mathbf{y}$ by using PLS, PCR, or TR for $\boldsymbol{\alpha}^{[k]}$, where $\mathbf{\Phi}^{[k]} = \mathbf{X}\mathbf{F}^{[k]}$
3	Recover $\mathbf{a}^{[k]}$ by using back-substitution $\mathbf{a}^{[k]} = \mathbf{F}^{[k]}\boldsymbol{\alpha}^{[k]}$
4	Compute primal regression vector via the relation $\mathbf{b}^{[k]} = \mathbf{X}^{[T]}\mathbf{a}^{[k]}$
5	Set $k = k + 1$ and go to Step 1

to the data before the current data instantiation is retired and replaced with a new one. Moreover, with such large volumes of spectra, one will also need to be vigilant about the increased likelihood of suspect data (e.g., inaccurate or missing data). The reliability of sparse methods is moot if the imprecision of the input limits the resolution of the output. As a result, it might be more prudent to opt for fast, robust, or incremental methods, as opposed to sparse ones (ideally, we would like to combine all of these traits).

We might also opt for data compression methods that find meaningful subsets of existing samples and wavelengths prior to regression or classification. In this context, subset selection algorithms deserve attention. For example, non–subset selection algorithms such as PCA and PLS construct low-rank approximations of the spectra. However, what we gain in compression, we lose in interpretation. The score and loading vectors used in the approximation are meta-features, i.e., they are linear combinations of all the wavelengths and samples, respectively. Instead of using linear combinations involving all rows or columns, there are algorithms, e.g., CUR matrix decompositions[77] and rank-revealing QR factorizations,[78,79] which construct low-rank approximations by using a small number of the original rows (samples) and/or columns (wavelengths). In this case, interpretation and highly accurate compression are simultaneously achieved. After a subset selection algorithm is applied to the spectra, a sparse method can then be applied on the reduced "original-feature" data set.

ACKNOWLEGMENT

The authors gratefully acknowledge InLight Solutions, Inc., for making the blood data set available to us for analysis.

1. L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, S. B. Engelsen. "Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy". Appl. Spectrosc. 2000. 54(3): 413-419.
2. R. Leardi. "Genetic Algorithms in Chemometrics and Chemistry: A Review". J. Chemom. 2001. 15(7): 559-569.
3. D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. de Noord. "Genetic Algorithms as a Tool for Wavelength Selection in Multivariate Calibration". Anal. Chem. 1995. 67(23): 4295-4301.
4. T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim. "Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles". Anal. Chem. 2009. 81(7): 2581-2590.
5. I.G. Chong, C.H. Jun. "Performance on Some Variable Selection Methods when Multicollinearity is Present". Chemometr. Intell. Lab. Syst. 2005. 78(1-2): 103-112.
6. W.J. Fu. "Penalized Regressions: The Bridge Versus the LASSO". J. Comput. Graph Stat. 1998. 7(3): 397-416.
7. B. Efron, I. Johnstone, T. Hastie, R. Tibshirani. "Least-Angle Regression". Ann. Stat. 2004. 32(2): 407-499.
8. J. Friedman, T. Hastie, H. Hölfling, R. Tibshirani. "Pathwise Coordinate Optimization". Ann. App. Stat. 2007. 1(2): 302-332.
9. Standford University "SparseLab: Seeking Solutions to Linear Solutions of Equations". 2007. http://sparselab.stanford.edu [accessed Jan 29, 2013].
10. J. Liu, S. Ji, J. Ye. Arizona State University. "SLEP: Sparse Learning with Efficient Projections". 2009. http://www.public.asu.edu/~jye02/Software/SLEP [accessed Jan 29, 2013].
11. M. Schmidt, G. Fung, R. Rosales. "Fast Optimization Methods for L_1 Regularization: A Comparative Study and Two New Approaches". Paper presented at: European Conference on Machine Learning. Warsaw, Poland: Sept. 17–21, 2007.
12. M. Mørup, L. Clemmensen. "Multiplicative Updates for the LASSO". Paper presented at: IEEE International Workshop on Machine Learning for Signal Processing. Thessaloniki, Greece: Aug. 27–29, 2007.
13. A. Hoerl. "Application of Ridge Analysis to Regression Problems". Chem. Eng. Prog. 1962. 58: 54-59.
14. A. Tikhonov. "Solution of Incorrectly Formulated Problems and the Regularization Method". Dokl. Akad. Nauk. SSSR. 1963. 151: 501-504.
15. J. Claerbout, F. Muir. "Robust Modeling of Erratic Data". Geophysics. 1973. 38(5): 826-844.
16. H.L. Taylor, S.C. Banks, J.F. McCoy. "Deconvolution with the l_1 Norm". Geophysics. 1979. 44(1): 39-52.
17. S. Levy, P. Fullagar. "Reconstruction of a Sparse Spike Train from a Portion of Its Spectrum and Application to High-Resolution Deconvolution". Geophysics. 1981. 46(9): 1235-1243.
18. F. Santosa, W.W. Symes. "Linear Inversion of Band-Limited Reflection Seismograms". SIAM J. Sci. Stat. Comp. 1986. 7(4): 1250-1254.
19. R.J. Tibshirani. "Regression and Shrinkage and Selection via the LASSO". J. Royal Statist. Soc. Ser. B. 1996. 58(1): 267-288.
20. H. Zou. "The Adaptive LASSO and Its Oracle Properties". J. Am. Stat. Assoc. 2006. 101(476): 1418-1429.
21. H. Zou, T. Hastie. "Regularization and Variable Selection via the Elastic Net". J Royal Statist. Soc. Ser. B. 2005. 67(2): 301-320.
22. P.C. Hansen. Rank-deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion. Philadelphia, PA: SIAM Press, 1998.
23. J.H. Kalivas. "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(6): 218-230.
24. H. Akaike. "A new look at the statistical model identification". IEEE Transact. Auto. Control. 1974. 19(6): 716-723.
25. G.E. Schwarz. "Estimating the Dimension of a Model". Ann. Stat. 1978. 6(2): 461-464.
26. D.M. Haaland, E.V. Thomas. "Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information". Anal. Chem. 1988. 60(11): 1193-2002.
27. T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning: Data Mining, Prediction and Inference. New York, NY: Springer, 2009. 2nd ed.
28. M.R. Osborne, B. Presnell, B.A. Turlach. "On the LASSO and Its Dual". J. Comput. Graph Stat. 2000. 9(2): 319-337.
29. P.J. de Groot, H. Swierenga, G.J. Postma, W.J. Melssen, L.M.C. Buydens. "Effect on the Partial Least-Squares Prediction of Yarn Properties Combining Raman and Infrared Measurements and Applying Wavelength Selection". Appl. Spectrosc. 2003. 57(6): 642-648.
30. H. Swierenga, F. Wülfert, O.E. de Noord, A.P. de Weijer, A.K. Smilde, L.M.C. Buydens. "Development of Robust Calibration Models in Near Infra-Red Spectrometric Applications". Anal. Chim. Acta. 2000. 411(1, 2): 121-135.
31. D. Özdemir, R. Williams. "Multi-instrument Calibration in UV–Visible Spectroscopy Using Genetic Regression". Appl. Spectrosc. 1999. 53(2): 210-217.
32. F.A. Honorato, R.K.H. Galvão, M.F. Pimentel, B. de Barros Neto, M.C.U. Araújo, F.R. de Carvalho. "Robust Modeling for Multivariate Calibration Transfer by the Successive Projections Algorithm". Chemom. Intell. Lab. Syst. 2005. 76(1): 65-72.
33. J.H. Kalivas, G.S. Siano, E. Andries, H.C. Goicoechea. "Tikhonov Regularization Approaches for Calibration Maintenance and Transfer". Appl. Spectrosc. 2009. 63(7): 800-809.
34. M.R. Kunz, J. Ottaway, J.H. Kalivas, E. Andries. "Impact of Standardization Sample Design on Tikhonov Regularization Variants for Spectroscopic Calibration Maintenance and Transfer". J. Chemom. 2010. 24(3-4): 218-229.
35. M.R. Kunz, J.H. Kalivas, E. Andries. "Model Updating for Spectral Calibration Maintenance and Transfer Using 1-Norm Variants of Tikhonov Regularization". Anal. Chem. 2010. 82(9): 3642-3649.
36. D.M. Haaland, D.K. Melgaard. "New Augmented Classical Least Squares for Improved

Quantitative Spectral Analyses". Vib. Spectrosc. 2002. 29(1): 171-175.
37. D.M. Haaland, D.K. Melgaard. "New Prediction Augmented Classical Least Squares (PACLS): Application to Unmodeled Interferents". Appl. Spectrosc. 2000. 54(9): 1303-1312.
38. D.M. Haaland, D.K. Melgaard. "New Classical Least Squares–Partial Least-Squares Hybrid Algorithm for Spectral Analyses". Appl. Spectrosc. 2001. 55(1): 1-8.
39. D.K. Melgaard, D.M. Haaland, C.M. Wehlburg. "Concentration Residual Augmented Classical Least Squares (CRACLS): A Multivariate Calibration Method with Advantages over Partial Least Squares". Appl. Spectrosc. 2002. 56(5): 615-624.
40. C.M. Wehlburg, D.M. Haaland, D.K. Melgaard, L.E. Martin. "New Hybrid Algorithm for Maintaining Multivariate Quantitative Calibrations of a Near-Infrared Spectrometer". Appl. Spectrosc. 2002. 56(5): 605-614.
41. R. Manne. "Analysis of Two Partial-Least-Squares Algorithms for Multivariate Calibration". Chemom. Intell. Lab. Syst. 1987. 2(1-3): 187-197.
42. H. Zou, T. Hastie, R. Tibshirani. "Sparse Principal Component Analysis". J. Comput. Graph Stat. 2004. 15(2): 262-286.
43. N.T. Trendafilov, I.T. Joliffe. "Projected Gradient Approach to the Numerical Solution of the SCoTLASS". Comput. Stat. Data. An. 2006. 50(1): 242-253.
44. A. D'Aspremont, L. Ghaoui, M.I. Jordan, G.R.G. Lanckriet. "A Direct Formulation for Sparse PCA Using Semidefinite Programming". SIAM Rev. 2007. 49(3): 434-448.
45. H. Shen, J.H. Huang. "Sparse Principal Component Analysis via Regularized Low-Rank Matrix Approximation". J. Multivariate Anal. 2008. 99(6): 1015-1034.
46. M. Journée, Y. Nesterov, P. Richtárik, R. Sepulchre. "Generalized Power Method for Sparse Principal Component Analysis". J. Mach. Learn. Res. 2010. 11: 517-553.
47. G.H. Fu, Q.S. Xu, H.D. Li, D.S. Cao, Y.Z. Liang. "Elastic Net Grouping Variable Selection Combined with Partial Least-Squares Regression (EN-PLSR) for the Analysis of Strongly Multi-collinear Spectroscopic Data". Appl. Spectrosc. 2011. 65(4): 402-408.
48. H. Chun, S. Keles. "Sparse Partial Least-Squares Regression for Simultaneous Dimension Reduction and Variable Selection". J. Royal. Statist. Soc. Ser. B. 2010. 72(1): 3-25.
49. K.A. Lê Cao, D. Rossouw, C. Rabert-Granié, P. Besse. "A Sparse PLS for Variable Selection when Integrating OMICs Data". Stat. Appl. Genet. Mol. Biol. 2008. 7: Article 35: 1-29.
50. M. Yuan, Y. Lin. "Model Selection and Estimation in Regression with Grouped Variables". J. Royal Statist. Soc. Ser. B. 2005. 68(1): 49-67.
51. M. Schmidt, G. Fung, R. Rosales. "Optimization Methods for L_1 Regularization". University of British Columbia Technical Report. 2009. TR-2009-19.
52. A. Tarantola. Inverse Problem Theory and Model Parameter Estimation. Philadephia, PA: SIAM Press, 2005.
53. J. Ottaway, J.H. Kalivas, E. Andries. "Spectral Multivariate Calibration with Wavelength Selection Using Variants of Tikhonov Regularization". Appl. Spectrosc. 2009. 64(12): 1388-1395.
54. E. Andries. "Sparse Models by Iteratively Reweighted Feature Scaling: A Framework for Wavelength and Sample Selection". J. Chemom. 2013. doi: 10.1002/cem.2492.
55. I. Gorodnitsky, B. Rao. "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum-Norm Algorithm". IEEE T Signal Proces. 1997. 45(3): 600-615.
56. B. Rao, K. Kreutz-Delgado. "An Affine Scaling Methodology for Best Basis Selection". IEEE T Signal Process. 1999. 47(1): 187-200.
57. E.J. Candès, M. Wakin, S. Boyd. "Enhancing Sparsity by Reweighted l_1 Minimization". J. Fourier Anal. Appl. 2007. 14(5): 877-905.
58. E.J. Candès., M. Wakin. "An Introduction to Compressive Sampling". IEEE Signal Process. Mag. 2008. March: 21-30.
59. Rice University. "Compressive Sensing Resources. References and Software. 2012. Digital Signal Processing Group. http://dsp.rice.edu/cs [accessed Jan 29, 2013].
60. H. Li, Y. Liang, Q. Xu, D. Cao. "Key Wavelengths Screening Using Competitive Adaptive Reweighted Sampling Method for Multivariate Calibration". Anal. Chim. Acta. 2009. 648(1): 77-84.
61. A. Kondylis, J. Whittaker. "Adaptively Preconditioned Krylov Spaces to Identify Irrelevant Predictors". Chemom. Intell. Lab. Syst. 2010. 104(2): 205-213.
62. E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, M. Tygert. "Randomized Algorithms for the Low-Rank Approximation of Matrices". P. Natl. Acad. Sci. USA. 2007. 104(51): 20167-20172.
63. V. Rokhlin, A. Szlam, M. Tygert. "A Randomized Algorithm for Principal Component Analysis". SIAM J. Matrix Anal. A. 2009. 31(3): 1100-1124.
64. N. Halko, P.-G. Martinsson, Y. Shkolnisky, M. Tygert. "An Algorithm for the Principal Component Analysis of Large Data Sets". SIAM J. Sci. Comput. 2011. 33(5): 2580-2594.
65. N. Halko, P.-G. Martinsson, J.A. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". SIAM Rev. 2011. 53(2): 217-288.
66. N. Cristianini, J. Shawe-Taylor. An Introduction to Support-Vector Machines and Other Kernel-Based Learning Methods. Cambridge, UK: Cambridge University Press, 1999.
67. J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle. Least-Squares Support-Vector Machines. Singapore: World Scientific Pub. Co., 2002.
68. M. Welling. University of California at Irvine. "Kernel Support-Vector Regression: Max Welling's Classnotes in Machine Learning". 2009. http://www.ics.uci.edu/~welling/classnotes/classnotes.html [accessed Jan 29, 2013].
69. J.N. Franklin. "Minimum Principles for Ill-Posed Problems". SIAM J. Math. Anal. 1978. 9(4): 638-650.
70. Eigenvector Research. "NIR of Corn Samples for Standardization Benchmarking". 2005. http://www.eigenvector.com/data/Corn/index.html [accessed Jan 29, 2013].
71. P.C. Williams., P.D.K. Grain. "Wheat Functionality as Measured by Diffuse Reflectance of Whole-Grain Canadian Wheat". International Diffuse Reflectance Conference. 2008. http://www.idrc-chambersburg.org/ss20082012.html [accessed Jan 29, 2013].
72. D. Abookasis, J.J. Workman. "Application of Spectra Cross-correlation for Type II Outliers Screening During Multivariate Near-infrared Spectroscopic Analysis of Whole Blood". Chemometr Intell. Lab. 2011. 107(2): 303-311.
73. K. Sjöstrand, L. Clemmensen, R. Larsen, B. Ersbøll. "SpaSM: a MATLAB Toolbox for Performing Sparse Regression". 2012. http://www2.imm.dtu.dk/projects/spasm/ [accessed Jan 29, 2013].
74. W. Clarke, D. Cox, L. Gonder-Frederick, W. Carter, S.L. Pohl. "Evaluating Clinical Accuracy of Systems for Self-Monitoring of Blood Glucose". Diabetes Care. 1987. 10: 622-628.
75. G. Gung, O.L. Mangasarian. "Proximal Support-Vector Classifiers". University of Wisconsin. Data Mining Institute Technical Report. 2001. TR-01-02.
76. L. Clemmensen, T. Hastie, D. Witten, B. Ersbøll. "Sparse Discriminant Analysis". Technometrics. 2011. 53(4): 406-413.
77. M.W. Mahoney, P. Drineas. "CUR Matrix Decompositions for Improved Data Analysis". Proc. Natl. Acad. Sci. USA. 2009. 106(3): 697-702.
78. T.F. Chan, P.C. Hansen. "Low-Rank Revealing QR Factorizations". Numer. Lin. Algebra Appl. 1994. 1(1): 33-44.
79. M. Gu, S.C. Eisenstat. "Efficient Algorithms for Computing a Strong Rank-Revealing QR Factorization". SIAM J. Sci. Comput. 1996. 17(4): 848-869.