Research article

Journal of CHEMOMETRICS

Received: 21 February 2015, Revised: 28 August 2015, Accepted: 31 August 2015, Published online in Wiley Online Library: 5 October 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2761

Model selection for partial least squares calibration and implications for analysis of atmospheric organic aerosol samples with mid-infrared spectroscopy

Satoshi Takahama[a]* and Ann M. Dillner[b]

In developing partial least squares calibration models, selecting the number of latent variables used for their construction to minimize both model bias and model variance remains a challenge. Several metrics exist for incorporating these trade-offs, but the cost of model parsimony and the potential for underfitting on achievable prediction errors are difficult to anticipate. We propose a metric that penalizes growing model variance against decreasing bias as additional latent variables are added. The magnitude of the penalty is scaled by a user-defined parameter that is formulated to provide a constraint on the fractional increase in root mean square error of cross-validation (RMSECV) when selecting a parsimonious model over the conventional minimum RMSECV solution. We evaluate this approach for quantification of four organic functional groups using 238 laboratory standards and 750 complex atmospheric organic aerosol mixtures with mid-infrared spectroscopy. Parametric variation of this penalty demonstrates that increase in prediction errors due to underfitting is bounded by the magnitude of the penalty for samples similar to laboratory standards used for model training and validation. Imposing an ensemble of penalties corresponding to a 0–30% allowable increase in RMSECV through sum of ranking differences leads to the selection of a model that increases the actual RMSECV up to 20% for laboratory standards but achieves an 85% reduction in the mean error in predicted concentrations for environmental mixtures. Partial least squares models developed with laboratory mixtures can provide useful predictions in complex environmental samples, but may benefit from protection against overfitting. © 2015 The Authors. *Journal of Chemometrics* published by John Wiley & Sons Ltd.

Additional supporting information may be found in the online version of this article at the publisher's web site

Keywords: multivariate calibration; partial least squares (PLS); bias/variance tradeoff; over-fitting; latent variable

1. INTRODUCTION

When the relationship between measured instrumental signals and response variables becomes difficult to decode, statistical methods of multivariate calibration can be used to develop models for quantitative prediction [1]. This challenge is particularly salient for characterization of organic functional group abundances in atmospheric aerosol samples (containing an ensemble of particles within a designated size range) by mid-infrared (MIR) spectroscopy [2]. These complex mixtures consisting of tens to hundreds of thousands of different types of organic molecules originate from the combination of directly emitted compounds and products of atmospheric photooxidation in the gas and condensed phases and pose challenges for characterization by any measurement technique [e.g., Hallquist *et al.* 3]. For Fourier transform infrared spectroscopy (FTIR) analysis, the challenge is manifested in broad absorption profiles originating from strong overlap of contributions from various functional groups present in the sample.

Despite the complex structure of the spectroscopic signal, MIR spectra of such samples contain a wealth of information and have been used to infer contributions from various source classes and extent of atmospheric processing of organic aerosol composition [4–6]. Estimation of functional group abundances in these samples is the first step in reconstructing estimates of the total organic aerosol burden and is relevant for interpreting the contribution from emission sources and atmospheric chemistry to a geographical location or region. In such applications, partial least squares (PLS) regression [1,7] has been used to develop models for quantitative calibration [8–11]. A necessary assumption is that such models developed from simpler standards prepared

* *Correspondence to: S. Takahama, ENAC/IIESwiss, Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland*
E-mail: satoshi.takahama@epfl.ch

[a] *S. Takahama*
ENAC/IIE, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland

[b] *A. M. Dillner*
University of California-Davis, Davis, CA, USA

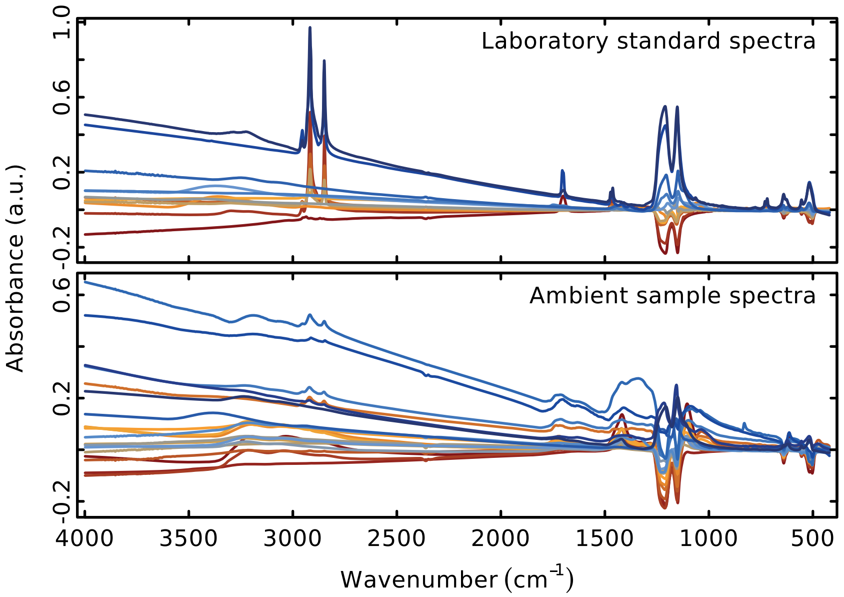

Figure 1. Example absorbance spectra for 12 laboratory standards and 20 ambient samples. For this illustration, an averaged spectrum of 54 blank samples have been subtracted to reduce the influence of scattering and absorption (still visible as strongly sloping background and interference near 1200 cm^{-1}) from the polytetrfluoroethylene substrates used in sample collection.

in the laboratory, where reference values are known and can be extrapolated to these complex samples that differ significantly in composition (i.e., number and types of molecules) and spectral structure.

In cases where laboratory standards used for calibration and environmental samples differ significantly (illustrated in Figure 1), it is conceivable that the effects of model misspecification can be amplified. Minimizing the prediction error against reference samples in the form of root mean square error (RMSE) is commonly used as an objective for model selection [12]. However, the true predictive performance of a model for new samples can be overestimated in that the RMSE metrics capture variations in bias but do not adequately account for the growth of variance as a function of model complexity, leading to selection of overly fitted PLS models containing more than the necessary number of latent variables (LVs) [13,14].

Researchers have previously proposed attainment of parsimonious models through consideration of the bias–variance trade-off in various forms. Alongside bias measures (e.g., RMSE calculated against various validation samples), model complexity or variance has been characterized in the form of effective rank [15,16], pseudo degrees of freedom [17,18], or some property of the regression vector (coefficients), for example, its two-norm magnitude [14,19] or 'jaggedness' introduced by oscillations [20,21]. These opposing measures have been evaluated along a Pareto curve [22,23] or combined together in a single metric [e.g., 14,18,21,24].

In this manuscript, we address the implications of model misspecification on prediction errors of laboratory standards and complex environmental mixtures, and methods for its prevention. We revisit the development of PLS calibration models for the quantification of functional groups in ambient aerosol samples previously described by [11] and explore the sensitivity of model selection on the formulation of metrics targeting parsimony. In this objective, we propose a modification of a metric described by [21] and [18]. We introduce a parameter to scale an arbitrary indicator of variance and penalize its growth against a measure of decreasing bias. The magnitude of this penalty is defined relative to the minimum RMSE of cross-validation (CV), such that the cost of parsimony can be anticipated against the available estimate of achievable prediction error. We vary the penalty on the variance measure (magnitude computed from vector of regression coefficients is used in this work) to generate a large set of model performance curves and report on the resulting complexity and prediction errors for the selected models. Prediction errors and sensitivity of predicted concentrations to penalization are evaluated for functional groups in laboratory standards and in ambient samples with compositions lying outside of the mixture space of the calibration model. For ambient samples in which we lack true reference values, we additionally compare an aggregate estimate of organic carbon (OC) estimated by the sum of FTIR functional groups with the measurements of OC obtained by a different but widely used analytical technique. A set of models selected from an ensemble of model performance curves generated by our proposed metric and combined by sum of ranking differences (SRD) [25,26] are validated against an independent randomization test [27], and we further evaluate their suitability for application to laboratory and ambient samples.

2. METHODS

2.1. Multivariate calibration and model selection

2.1.0.1. Partial least squares. We use PLS implemented by the `pls` library [28] in the R statistical package [29] to estimate regression vectors $\hat{\boldsymbol{b}}$ for predicting univariate responses of functional group concentrations $\hat{\boldsymbol{y}}$ from a set of spectra arranged in a row-wise matrix $\boldsymbol{X}$:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}} \tag{1}$$

We use the NIPALS algorithm with 10-fold venetian blinds CV on calibration samples (Section 2.2) sorted according to the known concentration of the response variable. In this way, the distribution of concentrations in validation and test sets are arranged to be similar to each other in each permutation during CV.

2.1.0.2. Prediction error. Root mean square error is a common metric used to evaluate the difference between estimated ($\hat{\boldsymbol{y}}$) and observed values ($\boldsymbol{y}$) in the response variable:

$$RMSE_k = \sqrt{\frac{1}{N}\|\hat{\boldsymbol{y}}_k - \boldsymbol{y}\|^2} \tag{2}$$

$\|\cdot\|$ is used to denote the two-norm of a vector (also written as $\|\cdot\|_2$, but we omit the subscript by convention), and the subscript k indicates the value computed using k LVs. RMSE is often defined more specifically in terms of RMSEC, the RMSE of calibration defined by the overall fit of models to samples in the training set; RMSEV, the RMSE with respect to samples in a validation set; RMSECV, the RMSE of CV evaluated within the calibration set (comprising the training and validation sets); and RMSEP, the RMSE of prediction reserved for new samples used for evaluation (the 'test set'). In the conventional method of model selection, RMSECV is computed for models generated with $1, 2, \ldots, \kappa$ LVs, and the optimal number of LVs (k^*) is selected by a criterion of minimum RMSECV to establish the base case. Alternatively, RMSEV can be used in place of RMSECV to determine k^*; we use RMSECV and a value of $\kappa = 120$ in this work. We denote the solution corresponding to the minimum RMSECV as $RMSECV_{k^*} = \min_k\{RMSECV_k\}$. The mean error is another metric used to assess the prediction error and is defined as

$$Mean\ error_k = \frac{1}{N}\|\hat{\boldsymbol{y}}_k - \boldsymbol{y}\|_1 \tag{3}$$

$\|\cdot\|_1$ is used to denote the one-norm of a vector. While we do not use the mean error for model selection, we report its value alongside the RMSEP for model evaluation as it is a common metric used in the atmospheric modeling community [30].

2.1.0.3. Metrics. The metric outlined by Gowen *et al.* [21] and Kalivas and Palmer [18] weighs a scaled measure of bias (i.e., RMSE) against a scaled 'regression vector measure' (RVM) that characterizes the magnitude of variance according to some property of the regression vector (as we refer to the vector of regression coefficients in this work). This metric is denoted $M1$ in this manuscript and computed over $k = 1, 2, \ldots, \kappa$ models:

$$M1_k = \left(\frac{RMSE_k - \min_k\{RMSE_k\}}{\max_k\{RMSE_k\} - \min_k\{RMSE_k\}}\right) + \left(\frac{RVM_k - \min_k\{RVM_k\}}{\max_k\{RVM_k\} - \min_k\{RVM_k\}}\right) \tag{4}$$

where $\{x_k\}$ denotes the set of values for all models $\{x_k : k = 1, 2, \ldots, \kappa\}$. $M1$ is dependent on $\max_k\{RVM_k\}$, which effectively determines the largest number of LVs considered in the set of solutions because RVM generally increases with the number of LVs. Therefore, κ is considered to be a free parameter on which the metric is defined. Model selection according to $M1$ is more sensitive to κ than for the minimum RMSECV criterion (where the determination of k^* is insensitive to κ given that it is sufficiently high), and yet what value of κ to be used is not clear *a priori*. Furthermore, the effect on prediction error due to underfitting by the newly selected model is uncertain.

We therefore propose a reformulation of $M1$ that parameterizes the RVM penalty relative to the minimum RMSE value:

$$M2_k = \left(\frac{RMSE_k}{\min_k\{RMSE_k\}}\right) + \lambda\left(\frac{RVM_k - \min_k\{RVM_k\}}{\max_k\{RVM_k\} - \min_k\{RVM_k\}}\right) \tag{5}$$

For this metric, the minima and maxima are defined for the set $\{x_k : 1, 2, \ldots, \kappa = k^*\}$, where k^* is determined *a priori* by RMSECV or RMSEV. Using this metric, we define the number of LVs selected by this metric as $k^\dagger = \arg\min_k\{M2_k\}$. This formulation shares some similarities with the objective function to be minimized by ridge regression or canonical Tikhonov regularization [31,32], which can be written with regularization parameter η as

$$\mathcal{L}(\boldsymbol{b}) = \|\boldsymbol{X}\boldsymbol{b} - \hat{\boldsymbol{y}}\|^2 + \eta^2\|\boldsymbol{b}\|^2 = N \cdot RMSE^2 + \eta^2\|\boldsymbol{b}\|^2 \tag{6}$$

The essential commonality between $M2$ and $\mathcal{L}$ is the penalization of a fidelity term by an RVM scaled with a weighting coefficient. While $\boldsymbol{b}$ is found directly through $\min_{\boldsymbol{b}}\mathcal{L}$ in ridge regression (without projection onto LVs) according to the magnitude of η, we propose for PLS that $\boldsymbol{b}$ should be selected from a set of candidate solutions constructed from the $k^\dagger$ LVs determined by the magnitude of λ prescribed in $M2$.

The appealing property of the scaled formulation of $M2$ is that λ fixes the allowed RVM penalty as a fraction or factor of $\min_k\{RMSE_k\}$ and bound the increase on prediction error due to underfitting when a more parsimonious model is chosen. Furthermore, by defining $\lambda^* = \max_k\{RMSE_k\}/\min_k\{RMSE_k\} - 1$, we can bound the anticipated increase in $RMSE_{k^\dagger}$ with respect to the estimated magnitude of reduction in prediction errors achievable through incorporation of additional LVs (Section S1). Written in the notation defined earlier,

$$\frac{RMSE_{k^\dagger}}{\min_k\{RMSE_k\}} \leq M2_{k^\dagger} \leq 1 + \lambda \tag{7}$$

$$\frac{RMSE_{k^\dagger} - \min_k\{RMSE_k\}}{\max_k\{RMSE_k\} - \min_k\{RMSE_k\}} \leq \frac{\lambda}{\lambda^*} \tag{8}$$

Also, by specification of $\lambda = \lambda^*$, we obtain the same solution $k^\dagger$ that would be selected by using $M1$ when $\kappa = k^*$ (Section S1), while we obtain the conventional solution $k^\dagger = k^*$ when $\lambda = 0$.

The primary difference between $M1$ and $M2$ is that the latter defines the RVM penalty by scaling the structure of growth in RVM from 1 to k^*, rather than 1 to $\kappa \geq k^*$. It is unclear whether additional information contained in the RVM from k^* to κ is relevant for selecting a more parsimonious model containing fewer than k^* LVs; it is hoped that the benefits of defining a new metric that can bound the increase in RMSE due to underfitting will outweigh the cost of this omission. Model selection by $M2$ in this way closely follows another heuristic of accepting an alternate solution corresponding to a fixed increase in RMSE (e.g., 10%, [33]) but additionally considers the dependence of model complexity and increased prediction variance on the number of LVs. Physically plausible values of λ may possibly be reasoned out based on an estimated uncertainty in RMSECV or RMSEV. In this work, we vary λ parametrically from 0 to λ^* and examine predictions from the models selected.

2.1.0.4. Specification of root mean square error and regression vector measure. For both $M1$ and $M2$, any of RMSECV, RMSEV, or RMSEP can be used. Using $M1$, Gowen *et al.* [21] found that using RMSECV for a fixed value of $\kappa = 20$ indicated instances of underfitting for their data set; [18] suggests that similar results are obtained with either metric. Using RMSECV can result in a value of $k^\dagger$ that is systematically less than or equal to the value selected in the scenario where RMSEC is used, but can be compensated by the selection of a larger κ (for $M1$) or smaller λ (for $M2$). To facilitate comparisons across all metrics, in this work we specify RMSECV as the bias measures of $M1$ and $M2$ to be consistent with the determination of $k^* = \arg\min_k\{RMSECV_k\}$ as stated earlier.

As introduced in Section 1, the properties of a vector as embodied by RVM can be defined by magnitude ($\|\hat{\boldsymbol{b}}\|$), jaggedness ($\|\Delta\hat{\boldsymbol{b}}\|_1$), or Durbin–Watson statistic [18,20,21], among other characteristics. We choose to use $\|\hat{\boldsymbol{b}}\|$ as this magnitude has been shown to be proportional to the prediction variance [19] and related to the sensitivity and detection limit of the calibration model [14]. In our models, $\|\Delta\boldsymbol{b}\|_1$ increases monotonically with $\|\hat{\boldsymbol{b}}\|$ (Figure S4), so the main conclusions for this work are expected to be insensitive to this choice.

2.1.0.5. Ensemble and randomization approach to model selection. We further employ two techniques to assess the number of LVs for each of our calibration models. Given the uncertainty in selection of λ for $M2$, we adapt an ensemble scoring approach to consider the most suitable model according to a range of λ values together. SRD described by Héberger and co-workers [25,34,35] is a method recently applied in the context of model or parameter selection for PLS and Tikhonov regularization [26]. In this work, we use the modern MATLAB implementation of SRD provided by Kalivas *et al.* [26]. In SRD, each metric or 'merit' is rescored according to its respective ranking against the target solution, which we designate as the minimum value of $M2$ [26]. $M2$ is calculated for each fold of PLS CV (number of folds is $V = 10$ in this work), for a common number of LVs determined as the maximum value of k^* computed across all folds. This leads to a block or matrix with dimensions of $V \times k^*$ for a single value of λ, and λ is varied to generate multiple merit blocks. For validating SRD ranking results, we use V-fold CV with $V = 10$. We select as our solution the minimum value of the mean normalized SRD and designate the number of LVs as $k^‡$ to differentiate from $k^†$ used to denote the solution obtained with a single realization of λ used with $M2$. A paired Wilcoxon signed rank test between the $k = k^‡$ solution and all others can be performed to find an alternate number of LVs for which the normalized SRD scores are not statistically significantly different [26]. However, we find that this approach can undermine the constraint on the growth of RMSECV imposed by λ, so it is not used in this work.

A randomization test for PLS has been described by Wiklund *et al.* [27] and used in similar contexts of model selection [e.g., Gowen *et al.* 21]. We adapt the 'randtest' function provided by the `mdatools` package in R [36] for use with the centered NIPALS algorithm provided by the `pls` library [28] and calculate our test statistic from $P = 1000$ permutations for each component. In this test, the order of samples in the response or residual vector $\boldsymbol{y}$ is permuted P number of times, and the conditional test statistic (covariance between PLS scores and $\boldsymbol{y}$ vector obtained after extraction fitted components) is compared with its corresponding value obtained for the original response or residual vector without permutations [27]. The exceedances of the reference value over the P permutations are referred to as the 'overfitting risk' in this work and are expressed as a percentage. The value of this percentage is compared with a significance level by close analogy to p-values used in standard statistical tests [36]. The overfitting risk is estimated by the empirical cumulative probability distribution in `mdatools`, and the cumulative probability distribution is additionally calculated using the inverse Gaussian function fitted to the test statistic using the `statmod` [37] and `fitdistrplus` [38] libraries in R. As the overfitting risk estimated by the two methods were practically identical for our interpretation, we only present one value of the risk that corresponds to the empirical cumulative probability distribution estimate. Wiklund *et al.* [27] report that randomization tests on spectra without pre-treatment can result in extraction of model components that are not in order of decreasing relevance, effectively leading to observations of erratic estimates of the overfitting risk. As discussed in Section 2.2, the signal contribution from the substrate in comparison with analyte is substantial in our samples and is not removed with pre-treatment for this analysis. Therefore, we interpret the results from this randomization test only with a qualitative appreciation; to guard against erratic significance values, we determine the maximum number of LVs by finding a consecutive sequence of length two or greater for which the significance of the test statistic is less than or equal to 5% and 10% significance levels, and take the largest number of components from this sequence.

2.2. Data set

2.2.0.6. Mid-infrared spectra. The set of infrared spectra used in this work (examples shown in Figure 1) is particulate matter samples collected and analyzed on polytetrafluoroethylene filter substrates as previously described by Ruthenburg *et al.* [11] and Dillner and Takahama [39]. The set consists of 238 laboratory standards (single component to ternary mixture samples of sugars, dicarboxylic acids, an ester, and ketone compounds) and 744 ambient samples (collected from seven US Interagency Monitoring of PROtected Visual Environments (IMPROVE) monitoring network sites in 2011). Out of the 794 ambient samples originally available, 50 are excluded from this evaluation as Ruthenburg *et al.* [11] identified them as spectrally anomalous. MIR spectra consisting of 2784 wavenumbers spanning the range between 4000 and 420 cm^{-1} without background correction is used for this work [39]. Four functional groups are considered for quantification: alcohol COH (aCOH), carboxylic COH (cCOH), alkane CH (aCH), and carbonyl C=O (CO).

Ruthenburg *et al.* [11] used 2/3 of the laboratory samples ($n = 158$) for model development and the remaining 1/3 ($n = 80$) for model selection according to the minimum RMSEV criterion. In this work, we use the same 2/3 of laboratory standards for calibration with CV and leave the remaining 1/3 for evaluation of our capability to predict concentrations in similar laboratory samples (to contrast with predictions for ambient samples). The similarity in relative functional group abundances and the concentrations between calibration and test set samples are shown in Figure S5. The maximum number of LVs considered by Ruthenburg *et al.* [11] was less than that used for this work ($\kappa = 30$ instead of $\kappa = 120$) and resulted in the selection of different models (Table S1).

2.2.0.7. Thermal optical reflectance organic carbon. Organic carbon concentrations measured by thermal optical reflectance (TOR) analysis [40] are taken from the IMPROVE network database (http://views.cira.colostate.edu/fed/). These samples are collected on quartz fiber filters collocated with the polytetrafluoroethylene filter samples used for FTIR analysis. This technique analyzes the total carbon vaporized when filters are subjected to a temperature ramp under an inert and then oxidizing environment. The OC fraction of the total carbon (the balance being elemental carbon) is operationally defined according to monitored optical properties of the filter during the vaporization process [40].

2.2.0.8. Calculation of organic carbon from functional groups. We can estimate the OC content from infrared analysis by taking the inner product of the molar functional group concentrations $\boldsymbol{n} = [n_{\text{aCOH}}, n_{\text{cCOH}}, n_{\text{aCH}}, n_{\text{CO}}]$ with the average number of moles of carbon associated with each bond $\boldsymbol{c} = [n_{\text{C}}/n_{\text{aCOH}}, n_{\text{C}}/n_{\text{cCOH}}, n_{\text{C}}/n_{\text{aCH}}, n_{\text{C}}/n_{\text{CO}}]$, such that the mass of carbon is estimated as $\boldsymbol{c}^T\boldsymbol{n} \times 12.01$ g/mol. The values of $\boldsymbol{c}$ are not known precisely for ambient samples but are estimated based on typical molecular structures assumed to be present in the mixture and can be fractional quantities to prevent potential double counting of atoms [2,9,11,41,42]. For this work, we use values of $\boldsymbol{c} = [0, 0, .5, 1]$ as assumed by Ruthenburg *et al.* [11].

3. RESULTS AND DISCUSSION

First, we describe the similarity of U-shaped curves generated by *M*1 and *M*2 by varying their respective tuning parameters (Figure 2). Curves of RMSECV and $\|\hat{\boldsymbol{b}}\|$ used for calculation of *M*1 and *M*2 are also shown. $\|\hat{\boldsymbol{b}}\|$ is observed to increase monotonically for our models, while strict monotonicity is not observed for RMSECV. U-shaped curves are visible for *M*1 over the domain of $k = \{1, 2, \ldots, \kappa = 2k^*\}$, which we have specified for illustration. Unambiguous U-shaped curves within the domain of $k = \{1, 2, \ldots, k^*\}$ are generally observed to emerge for *M*2 when the value of λ is approximately greater than one ($\lambda = 0.2$ and $\lambda = 5.0$ are illustrated). One consequence of this pattern is that for low values of λ, $1+\lambda$ is a close approximation of $RMSECV_{k^\dagger}/RMSECV_{k^*}$. As λ is increased, the actual increase in $RMSECV_{k^\dagger}/RMSECV_{k^*}$ will become increasingly small compared with that of $1 + \lambda$. For this data set, the solution obtained with *M*1 for $\kappa \leq 2k^*$ corresponds to a penalty for *M*2 of $\lambda > 0.2$. Further comparisons of the two metrics are shown in Section S2, in which we also summarize that the selected model is sensitive to the specification of κ in *M*1, but how the variation in κ influences the increase in RMSECV of the selected solution is difficult to anticipate. Only solutions obtained using *M*2 and λ will be discussed in the following sections.

3.1. Evaluation of latent variable reduction on laboratory standards

The variation in $k^\dagger$ and the corresponding fit metrics as a function of λ as formulated by *M*2 is shown in Figure 3. We can see the clear trend of monotonic decrease in the number of LVs selected with increasing λ. In all cases, $RMSECV_{k^\dagger} \leq (1 + \lambda)RMSECV_{k^*}$, which is a property of the metric (Section 2.1). We do note that $1 + \lambda$ becomes an increasingly conservative upper bound of the actual increase in RMSECV over $RMSECV_{k^*}$ when λ is large.

We can see that a structure and magnitude similar to the variation in RMSECV with respect to λ are preserved in the

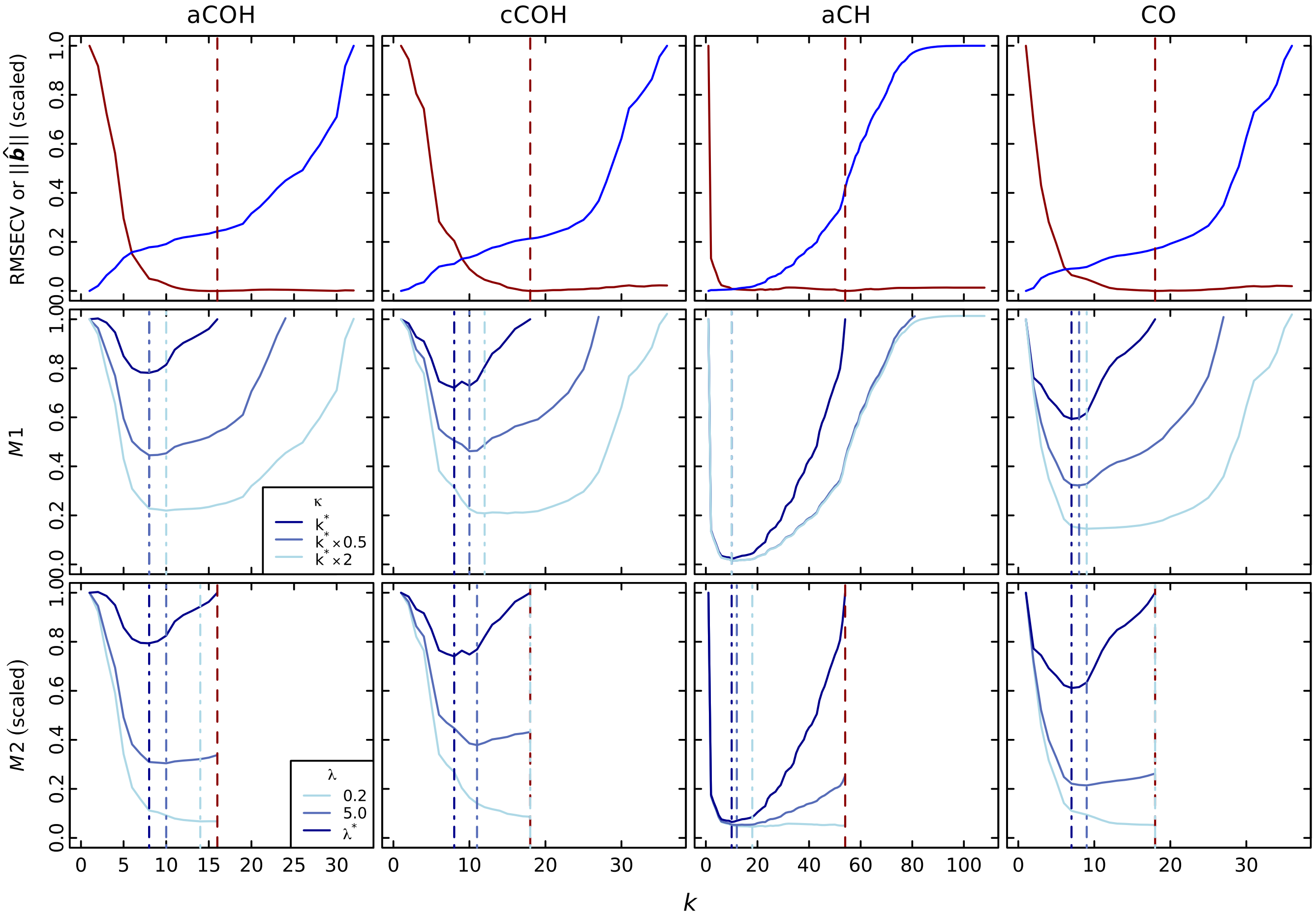

Figure 2. Root mean square error of cross-validation (RMSECV) (red), $\|\hat{\boldsymbol{b}}\|$ (blue), and example curves of *M*1 and *M*2 computed for calibration samples prepared in the laboratory for four organic functional groups considered in this work (shown in separate columns). For this figure, RMSECV and $\|\hat{\boldsymbol{b}}\|$ are offset by their minimum value and scaled by their range to lie within [0,1], and *M*2 is scaled by a factor $\min_k\{RMSECV_k\}/\max\{RMSECV_k\}$. For illustration, several arbitrary values of κ and λ have been selected for *M*1 and *M*2, respectively. Values of κ correspond to the nearest integer values of 1, 1.5, and 2 times k^*. Dark red vertical lines correspond to $k = k^*$, and vertical blue lines correspond to the values of $k = k^\dagger$. aCOH, alcohol COH; cCOH, carboxylic COH; aCH, alkane CH; CO, carbonyl C=O.

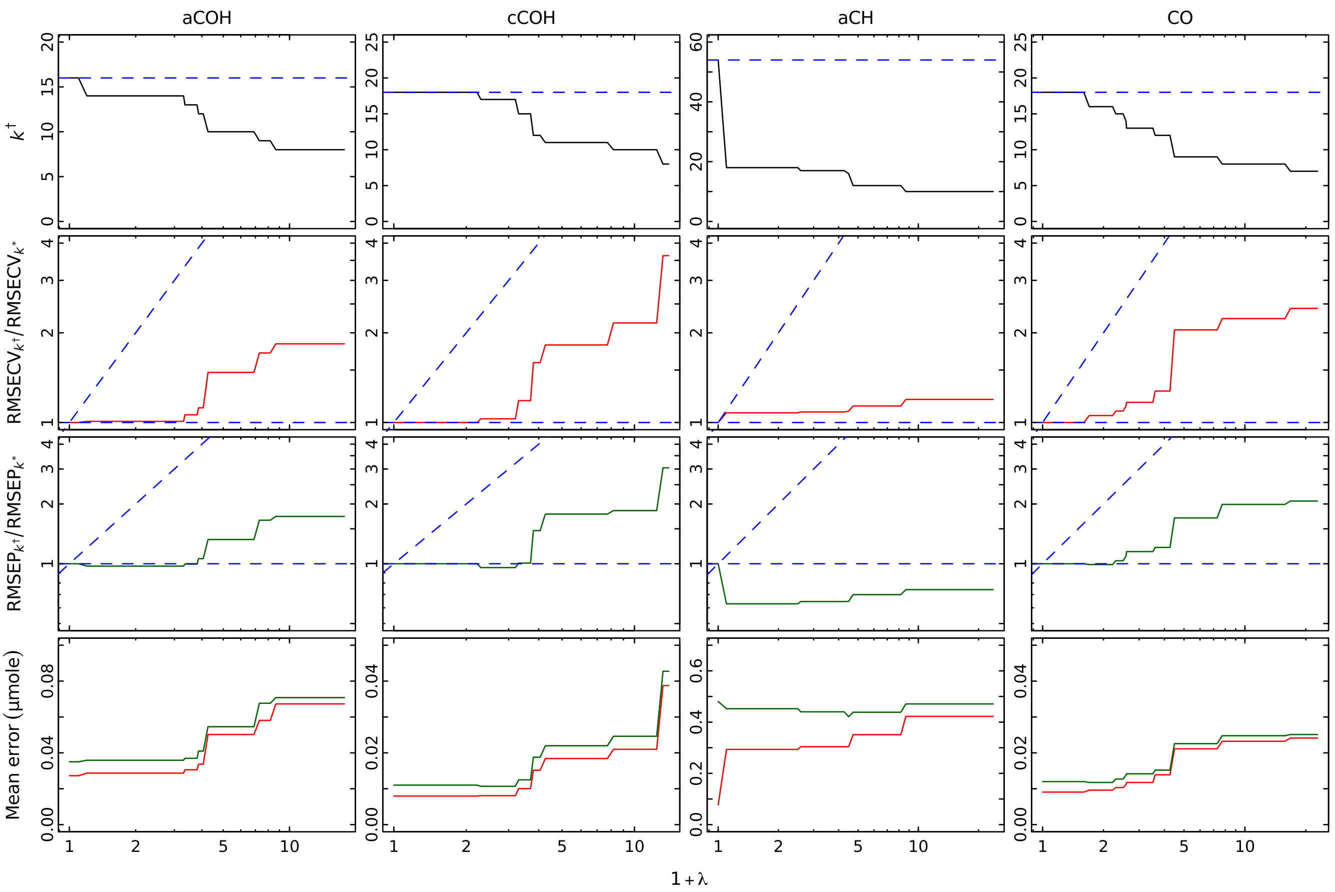

Figure 3. The number latent variables selected according to $k = k^\dagger$ (top panel) and the corresponding fit metrics for laboratory standards as the parameter λ is varied from 0 to $\lambda^* = \max\{RMSECV_k\}/\min_k\{RMSECV_k\} - 1$ for each variable (rows 2–4). Dotted blue horizontal lines correspond to the $k = k^*$ solution. Solid red lines are used to indicate evaluations for calibration samples, and solid green lines indicate evaluations for test set samples. Note that the *y*-axes for panels in rows 2 and 3 are in logarithmic scale and share the same limits across all columns. Blue diagonal lines in the second row of panels correspond to the $x = y$ line indicating the 'upper bound' for an increase relative to $RMSECV_{k^*}$ and correspond to $(1 + \lambda)$. The *x*-axes are also shown as $1 + \lambda$ in logarithmic scale. aCOH, alcohol COH; cCOH, carboxylic COH; aCH, alkane CH; CO, carbonyl C=O.

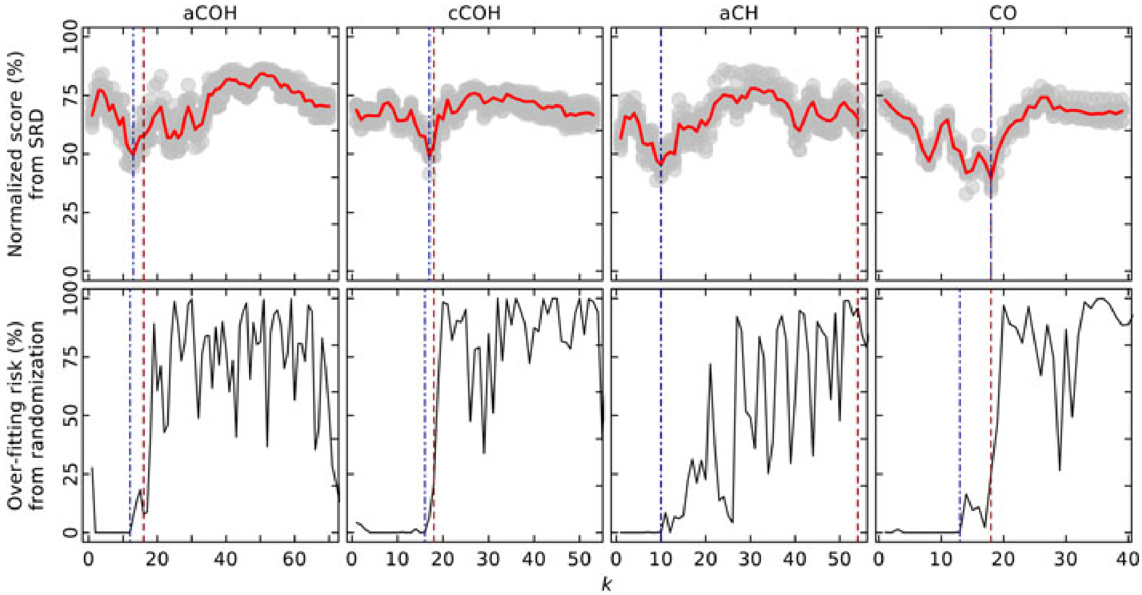

Figure 4. Top row: Normalized sum of ranking differences scores as a function of latent variables. Gray circles represent each realization of SRD cross-validation, and red lines indicate their means. Bottom row: Overfitting risk obtained from randomization test (Section 2.1). Dark red vertical lines indicate the model selected by the minimum root mean square error of cross-validation criterion ($k = k^*$), and blue vertical lines indicate the solution obtained by either SRD or randomization. aCOH, alcohol COH; cCOH, carboxylic COH; aCH, alkane CH; CO, carbonyl C=O.

corresponding RMSEP and mean error estimated for test set laboratory standards for most functional groups (Figure 3). The RMSEP of the test set is approximately an order of magnitude higher than RMSECV (a factor of 9 on average), and the RMSEP ranges between 0% and 305%. The difference in the mean error, however, remains small (28–34%) between the two sets of laboratory standards (Figures 3 and S2).

Alkane CH does not follow the pattern of variation in fitting statistics observed for the rest of the functional groups: The change in RMSEP and mean error is not anticipated by the variation in RMSECV. An RVM penalty of $\lambda = 0.1$ results in a reduction in the selected number of factors from 54 to 18, an increase in RMSECV of 8%, and a decrease in RMSEP of 40%. Over the range of λs explored, RMSECV changed the least for aCH compared with all other variables. While RMSECV increases by 20% for $\lambda = 23$, RMSEP does not increase above the value of $RMSEP_{k^*}$. The same conclusion is reached when random 10-fold CV is used (not shown), which possibly implicates the influence of structural differences between calibration and evaluation samples with respect to the absorption bands of aCH. Structural differences in absorption bands may be further reduced by rigorous statistical design of the calibration set when mixture composition of new samples can be anticipated, but model sensitivity to such differences can be problematic when predicting concentrations in more complex environmental samples (Section 3.3). The absorption bands of aCH (approximately 3000–2800 cm^{-1} [43]; Figure 1) in atmospheric aerosols are particularly feature rich, with absorbance patterns differing by hydrocarbon source type, for example, vegetative detritus [44] or various forms of fossil fuel combustion [6].

3.2. Model selection

The appropriate value of λ for this real data set is unknown, so we apply an ensemble modeling approach implemented with SRD. Normalized SRD scores calculated for $\lambda = \{0.0, 0.1, 0.2, 0.3\}$ are shown in Figure 4. The number of LVs selected is identical or similar to when $\lambda = \{0.0, 0.1, 0.2\}$, $\{0.1, 0.2, \ldots, 0.5\}$, and $\{0.0, 0.1, 0.2, \ldots, 1.0\}$ ensembles are used (Table I), with proximity to the $k = k^*$ solution with the exception of aCH (for which the number of LVs selected is lower by a factor of 5). While the upper bound of $1 + \lambda$ as defined in Equation 7 holds true for each fold of

Table I. Number of LVs selected according to different methods

Method	Number of LVs ($RMSECV_{k^\ddagger}/RMSECV_{k^*}$) aCOH	cCOH	aCH	CO
minimum RMSECV	16 (1.00)	18 (1.00)	54 (1.00)	18 (1.00)
SRD with $M2$; $\lambda = \{0.0, 0.1, 0.2\}$	11 (1.23)	17 (1.03)	12 (1.14)	14 (1.14)
SRD with $M2$; $\lambda = \{0.0, 0.1, 0.2, 0.3\}$	13 (1.06)	17 (1.03)	10 (1.20)	18 (1.00)
SRD with $M2$; $\lambda = \{0.0, 0.1, 0.2, \ldots, 0.5\}$	13 (1.06)	17 (1.03)	10 (1.20)	18 (1.00)
SRD with $M2$; $\lambda = \{0.0, 0.1, 0.2, \ldots, 1.0\}$	13 (1.06)	17 (1.03)	10 (1.20)	18 (1.00)
Randomization; 5% significance level	12 (1.12)	16 (1.11)	10 (1.20)	13 (1.17)
Randomization; 10% significance level	13 (1.06)	17 (1.03)	26 (1.13)	13 (1.17)

LVs, latent variables; aCOH, alcohol COH; cCOH, carboxylic COH; aCH, alkane CH; CO, carbonyl C=O; RMSECV, root mean square error of cross-validation; SRD, sum of ranking differences.

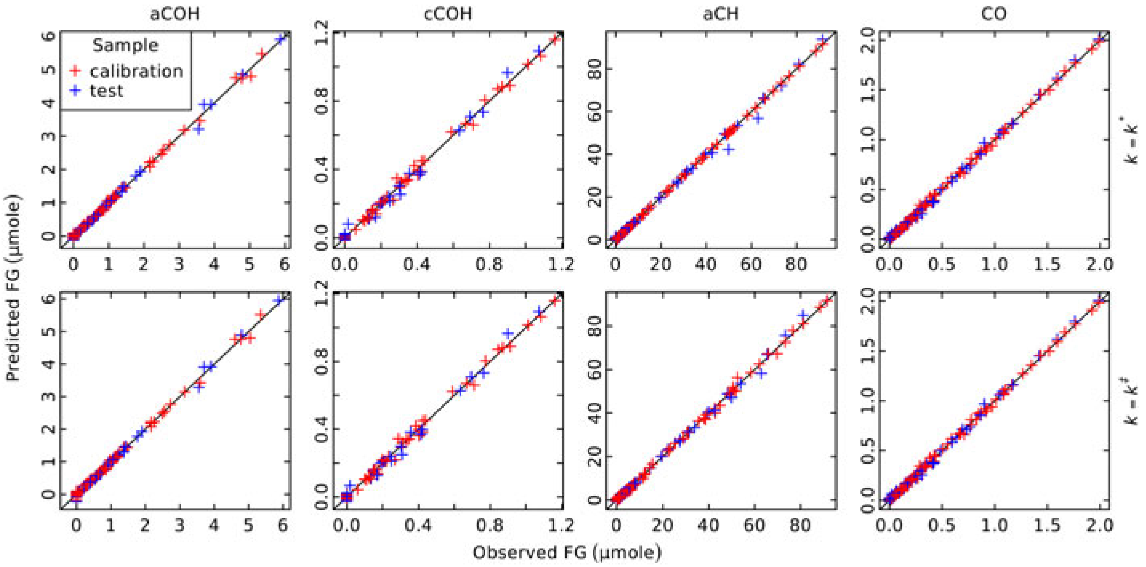

Figure 5. Comparison of predicted concentrations with observed concentrations of functional groups (FG) (shown across columns) in laboratory standards for solution selected by the minimum root mean square error of cross-validation criterion ($k = k^*$) and ensemble scoring ($k = k^\ddagger$) with $M2$ using multiple values of λ. aCOH, alcohol COH; cCOH, carboxylic COH; aCH, alkane CH; CO, carbonyl C=O.

CV, the relative magnitudes are rescored as rankings with respect to the RMSECV prior to aggregation. As a result, we note that the number of LVs selected does not monotonically decrease with the maximum λ of each ensemble and the bound of $1 + \lambda$ in the growth of RMSECV is not strictly obeyed when ensemble scoring is used (illustrated by the fact that the increase in RMSECV is 1.23 for aCOH when the maximum $1 + \lambda$ is 1.2). However, the ensemble scoring approach indicates that the $k^{\ddagger}$s selected are same or fewer than k^{*}s across functional groups (especially for aCH), and reasonable agreement can be found among the $k^{\ddagger}$s for different ensembles of λ.

We independently confirm by a randomization test at the 5% significance level that the number of relevant components is also less than or equal to the those calculated by the minimum RMSE criterion and is also similar to values calculated by SRD—especially for aCH (Table I). Figure 4 illustrates how the overfitting risk changes according to the number of LVs, and solutions selected at the 5% significance levels are also indicated. As shown in the same figure, the first component for aCOH exceeds 5%, but we attribute this anomaly to the presence of the substrate interference in the infrared spectra, and it is not considered for estimation of LVs. This is anticipated from the fact that pretreatment was not applied to remove the substrate interferences to the signal as described by Wiklund *et al.* [27] (Section 2.1). The solution for the 10% significance level does show a noticeable departure in the selected number of LVs for aCH compared with the other estimates, however indicating sensitivity to the choice of significance level combined with erratic variations in the overfitting risk.

For further analysis, we choose as our reference solution the number of LVs ($k = k^{\ddagger}$) determined by $M2$ with SRD for $\lambda = \{0.0, 0.1, 0.2, 0.3\}$. Figure 5 and increases in RMSECVs found in Table I show that the predicted concentrations in laboratory standards are insensitive to the selection of these solutions, but we discuss their appropriateness for application to samples outside of the domain of the calibration set in Section 3.3.

3.3. Implications for extrapolation

To evaluate implications for overfitting in environmental samples in which reference functional group concentrations are not available, we examine the changes in predicted concentrations of each variable for various values of λ and compare the aggregated estimates of FTIR OC with the measured TOR OC (Figure 6). We first note the similarity ($r > 0.8$) in predicted concentrations for each functional group between the k^{*} and $k^{\dagger}$ solutions for modest values of $\lambda < 0.2$, except for aCH. At $\lambda = 0.1$, the number of LVs

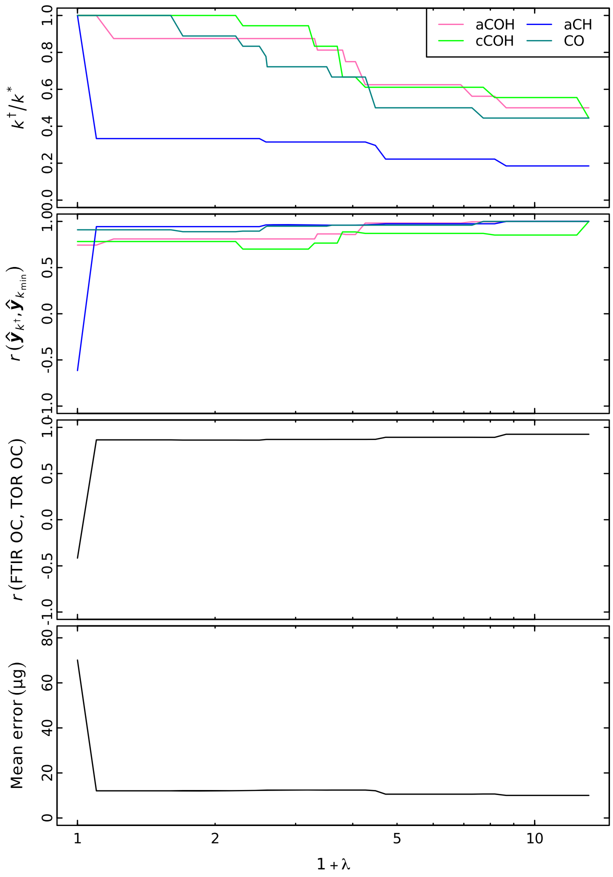

Figure 6. Evaluation of calibration models selected by various values of λ (shown on abscissa) applied to complex atmospheric aerosol mixtures. The number of latent variables relative to the minimum root mean square error of cross-validation solution ($k^{\dagger}/k^{*}$) determined for each functional group in the calibration models (developed from laboratory-generated mixtures) is shown in the top panel. The correlation coefficient of the estimated concentrations [$r(\hat{\mathbf{y}}_{k^{\dagger}}, \hat{\mathbf{y}}_{k_{\min}})$] with respect to the minimum latent variable solution in ambient samples is shown in the second panel. Summary statistics (mean error and Pearson's correlation coefficient, r) of Fourier transform infrared spectroscopy organic carbon (FTIR OC) estimates in atmospheric samples compared with reported values of collocated thermal optical reflectance (TOR) OC measurements as a function of the λ penalty imposed across all functional groups are shown in the bottom two panels. aCOH, alcohol COH; cCOH, carboxylic COH; aCH, alkane CH; CO, carbonyl C=O.

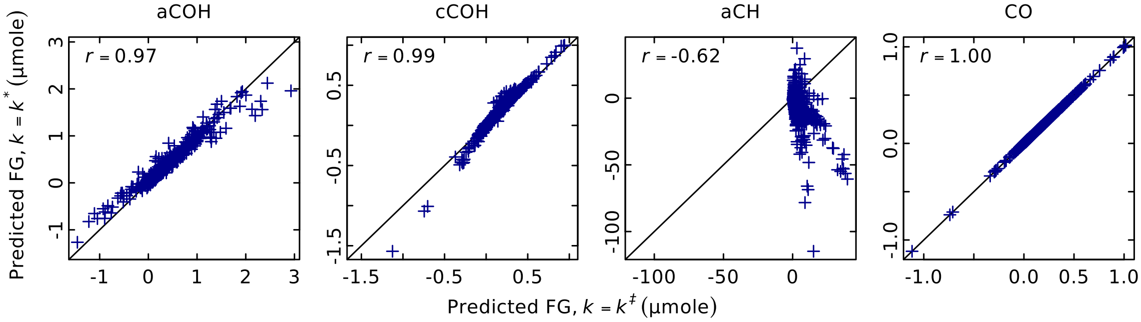

Figure 7. Comparison of functional group (FG) estimates in ambient samples for solutions selected by the minimum root mean square error of cross-validation criterion ($k = k^{*}$) and ensemble scoring ($k = k^{\ddagger}$) with $M2$ using multiple values of λ. Pearson's correlation coefficients (r) are shown in the upper left corner. aCOH, alcohol COH; cCOH, carboxylic COH; aCH, alkane CH; CO, carbonyl C=O.

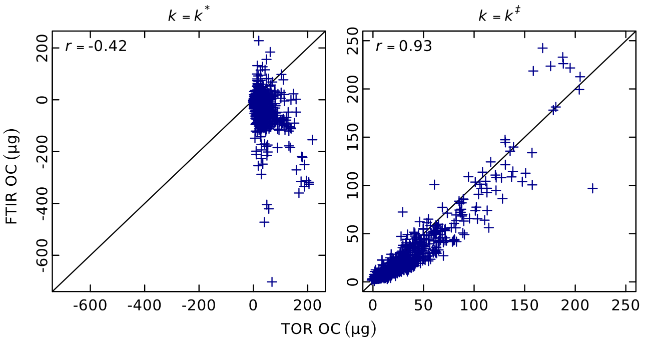

Figure 8. Comparison of estimated Fourier transform infrared spectroscopy organic carbon (FTIR OC) mass in ambient samples with collocated thermal optical reflectance (TOR) OC measurements using the model selected by the minimum root mean square error of cross-validation criterion ($k = k^*$) and ensemble scoring ($k = k^‡$) with *M*2 using multiple values of λ. Pearson's correlation coefficients (*r*) are shown in the upper left corner.

selected for aCH is reduced from 54 to 18, and this coincides with large differences in predicted concentrations (Figure 6, top two panels). For the selected solutions corresponding to $k = k^‡$, we also observe major differences ($r = -0.62$) with respect to the $k = k^*$ solution for aCH (Figure 7) when the number of LVs is reduced from 54 to 10. This change is highlighted in contrast to the insensitivity of predicted abundances for laboratory standards to the selected number of LVs (Section 3.2 and Figure 5).

Aggregating the OC content from a combination of the estimated functional groups and comparing with TOR OC, we find that a very large increase in the Pearson's correlation coefficient (from $r = -0.42$ to 0.86) and a reduction in the mean error (from 70 to 12 µg of OC mass on the filter) are observed for $\lambda = 0.1$, with continued agreement between the two estimates of OC with increasing λ (Figure 6, bottom two panels). For our selected value of $k^‡$s (Section 3.2), the error is further reduced to 10 µg (Figure 8). The agreement of FTIR OC with TOR OC is largely dictated by the variation in aCH, as aCH is estimated to compose 60–78% of OC mass in these samples. While separate sample collection and analytical artifacts exist for OC quantification by FTIR and TOR [45], we expect a general agreement between the two measurements and conclude that the $k^‡$ solution is more appropriate than the k^* solution.

A small difference in the correlation (from $r = 0.86$ to 0.93) between the two estimates of OC is observed between $\lambda = 0.1$ and greater values of λ (Figures 6) primarily because of the change in the number of factors selected for aCH, particularly for a certain class of spectra. While beyond the scope of this manuscript, focused study of such differences in sensitivity across ambient samples may further provide characterization of mixture composition for specific samples.

4. CONCLUSIONS

We propose a reformulation of a metric weighing bias and variance measures for model selection. The defining parameter, which frames the trade-off between model parsimony and the lowest prediction error, is changed from the total number of LVs considered to a penalization parameter interpreted as the permissible increase relative to the minimum RMSECV, for which we have better intuitive sense. We explore the impact of the parameter on model selection and estimation of organic functional group concentrations from infrared spectra. We build a calibration model from 158 laboratory samples and evaluate predictions for 80 laboratory samples similar to those in the calibration set and for 750 complex environmental mixtures in which true references are not available for calibration.

We find, expectedly, that the number of LVs selected for PLS generally decreases according to increasing penalization for larger RVM (used as an indication of model variance). For a number of models, we can predict concentrations in laboratory standards with modest variations in RMSECV (≤20%), but extension of these models to more complex mixtures leads to larger differences (greater than 100% in predicted concentrations). In comparison with an independent estimate of OC, we find that the model with a higher number of LVs as selected by the minimum RMSECV criterion is unable to estimate the mass of organic material in the samples.

As the appropriate choice of penalization parameter in our metric is not known, we use an ensemble scoring approach (SRD) to aggregate solutions for various penalty values. When using SRD, the actual increase in RMSECV for the selected solution can increase modestly above the maximum penalty specified but provides a consistent selection of LVs that is robust with respect to a range of penalty values considered and consistent with an independent randomization test. In our work, we demonstrate that the model selected by an ensemble of penalties corresponding to maximum allowable increase in RMSECV of 0%, 10%, 20%, and 30% yields an actual increase in RMSECV of 0–20% across functional groups. The same model drastically improves predictions in environmental samples, however reducing the mean error with respect to an independent metric of OC mass from 70 to 10 µg (an 85% reduction) and increasing the correlation between predictions and observations from $r = -0.42$ to $r = 0.93$.

As previously reported, PLS models selected on the basis of bias measure may be susceptible to overfitting, which becomes apparent when applying calibration models to more complex mixtures. In such an application, the conventional minimum RMSE metric may yield models that lead to gross errors. A reformulated metric combined with an ensemble scoring approach can provide some additional guidance for selecting a model that considers the cost of parsimony on increased prediction error, while guarding against larger errors incurred by overfitting.

Acknowledgements

The authors acknowledge the funding from EPFL and the IMPROVE program (National Park Service cooperative agreement P11AC91045) and the helpful comments from the anonymous peer reviewers.

REFERENCES

1. Martens H. *Multivariate Calibration*. John Wiley & Sons: New York, 1991.
2. Allen DT, Palen EJ, Haimov MI, Hering SV, Young JR. Fourier-transform infrared-spectroscopy of aerosol collected in a low-pressure impactor (LPI/FTIR) - method development and field calibration. *Aerosol Sci. Technol.* 1994; **21**(4): 325–342.
3. Hallquist M, Wenger JC, Baltensperger U, Rudich Y, Simpson D, Claeys M, Dommen J, Donahue NM, George C, Goldstein AH, Hamilton JF, Herrmann H, Hoffmann T, Iinuma Y, Jang M, Jenkin ME, Jimenez JL, Kiendler-Scharr A, Maenhaut W, McFiggans G, Mentel TF, Monod A, Prevot ASH, Seinfeld JH, Surratt JD, Szmigielski R, Wildt J. The formation, properties and impact of secondary organic aerosol: current and emerging issues. *Atmos. Chem. Phys.* 2009; **9**: 5155–5236.

4. Russell LM, Bahadur R, Ziemann PJ. Identifying organic aerosol sources by comparing functional group composition in chamber and atmospheric particles. *Proc. Natl. Acad. Sci. U.S.A.* 2011; **108**(9): 3516–3521.
5. Corrigan AL, Russell LM, Takahama S, Äijälä M, Ehn M, Junninen H, Rinne J, Petäjä T, Kulmala M, Vogel AL, Hoffmann T, Ebben CJ, Geiger FM, Chhabra P, Seinfeld JH, Worsnop DR, Song W, Auld J, Williams J. Biogenic and biomass burning organic aerosol in a Boreal forest at Hyytiälä, Finland, during HUMPPA-COPEC 2010. *Atmos. Chem. Phys.* 2013; **13**(24): 12233–12256.
6. Guzman-Morales J, Frossard A, Corrigan A, Russell L, Liu S, Takahama S, Taylor J, Allan J, Coe H, Zhao Y, Goldstein A. Estimated contributions of primary and secondary organic aerosol from fossil fuel combustion during the CalNex and Cal-Mex campaigns. *Atmos. Environ*. 2014; **88**: 330–340.
7. Wold S, Martens H, Wold H. The multivariate calibration-problem in chemistry solved by the PLS method. *Lecture Notes in Math.* 1983; **973**: 286–293.
8. Reff A, Turpin BJ, Porcja RJ, Giovennetti R, Cui W, Weisel CP, Zhang J, Kwon J, Alimokhtari S, Morandi M, Stock T, Maberti S, Colome S, Winer A, Shendell D, Jones J, Farrar C. Functional group characterization of indoor, outdoor, and personal PM2.5: results from RIOPA. *Indoor Air* 2005; **15**(1): 53–61.
9. Reff A, Turpin BJ, Offenberg JH, Weisel CP, Zhang J, Morandi M, Stock T, Colome S, Winer A. A functional group characterization of organic PM2.5 exposure: results from the RIOPA study RID C-3787-2009. *Atmos. Environ.* 2007; **41**(22): 4585–4598.
10. Coury C, Dillner AM. A method to quantify organic functional groups and inorganic compounds in ambient aerosols using attenuated total reflectance FTIR spectroscopy and multivariate chemometric techniques. *Atmos. Environ*. 2008; **42**(23): 5923–5932.
11. Ruthenburg TC, Perlin PC, Liu V, McDade CE, Dillner AM. Determination of organic matter and organic matter to organic carbon ratios by infrared spectroscopy with application to selected sites in the improve network. *Atmos. Environ*. 2014; **86**: 47–57.
12. Mevik BH, Cederkvist HR. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *J. Chemom.* 2004; **18**(9): 422–429.
13. Shao J. Linear model selection by cross-validation. *Amer. Statist. Assoc.* 1993; **88**(422): 486–494.
14. Green RL, Kalivas JH. Graphical diagnostics for regression model determinations with consideration of the bias/variance trade-off, vol. 60, September 18–20, 2000; 173–188.
15. Hansen P. *Rank-Deficient And Discrete Ill-Posed Problems*. Society for Industrial and Applied Mathematics: Philadelphia, 1998.
16. Seipel HA, Kalivas JH. Effective rank for multivariate calibration methods. *J. Chemom.* 2004; **18**(6): 306–311.
17. van der Voet H. Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. *J. Chemom.* 1999; **13**(3-4): 195–208.
18. Kalivas JH, Palmer J. Characterizing multivariate calibration tradeoffs (bias, variance, selectivity, and sensitivity) to select model tuning parameters. *J. Chemom.* 2014; **28**(5): 347–357.
19. Faber K, Kowalski BR. Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *J. Chemom.* 1997; **11**(3): 181–238.
20. Rutledge DN, Barros AS. Durbin Watson statistic as a morphological estimator of information content. *Anal. Chim. Acta* 2002; **454**(2): 277–295.
21. Gowen AA, Downey G, Esquerre C, O'Donnell CP. Preventing overfitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *J. Chemom.* 2011; **25**(7): 375–381.
22. Kalivas JH. Multivariate calibration, an overview. *Anal. Lett.* 2005; **38**(14): 2259–2279.
23. Stout F, Kalivas JH. Tikhonov regularization in standardized and general form for multivariate calibration with application towards removing unwanted spectral artifacts. *J. Chemom.* 2006; **20** (1-2): 22–33.
24. Höskuldsson A. The H-principle: new ideas, algorithms and methods in applied mathematics and statistics. *Chemom. Intell. Lab. Syst. Proceedings of the 3rd Scandinavian Symposium on Chemometrics (SSC3)* 1994; **23**: 1–28.
25. Héberger K, Kollár-Hunek K. Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *J. Chemom.* 2011; **25**(4): 151–158.
26. Kalivas JH, Héberger K, Andries E. Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods. *Anal. Chim. Acta* 2015; **869**(0): 21–33.
27. Wiklund S, Nilsson D, Eriksson L, Sjostrom M, Wold S, Faber K. A randomization test for PLS component selection. *J. Chemom.* 2007; **21**(10–11): 427–439.
28. Mevik B, Wehrens R. The PLS package: principal component and partial least squares regression in R. *J. Stat. Softw.* 2007; **18** (2): 1–24.
29. R Core Team. R: A language and Environment for Statistical Computing, R Foundation for Statistical Computing: Vienna, Austria, 2014.
30. Seinfeld JH, Pandis SN. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change (2nd edition). John Wiley & Sons: New York, 2006.
31. Tikhonov AN, Arsenin VI. *Solutions of Ill-Posed Problems*. Halsted Press: New York, 1977.
32. Kalivas JH. Overview of two-norm (l2) and one-norm (l1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance. *J. Chemom.* 2012; **26**(6): 218–230.
33. Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Series B-statistical Methodology* 2010; **72**: 3–25.
34. Héberger K. Sum of ranking differences compares methods or models fairly. *TrAC Trends in Anal. Chem.* 2010; **29**(1): 101–109.
35. Kollár-Hunek K, Héberger K. Method and model comparison by sum of ranking differences in cases of repeated observations (ties). *Chemom. Intell. Lab. Syst.* 2013; **127**: 139–146.
36. Kucheryavskiy S. *Mdatools: Multivariate Data Analysis for Chemometrics R Package Version 0.6.0*, 2015. http://CRAN.R-project.org/package=mdatools.
37. Smyth G, Hu Y, Dunn P, Phipson B, Chen Y. *Statmod: Statistical Modeling R Package Version 1.4.21*, 2015. http://CRAN.R-project.org/package=statmod.
38. Delignette-Muller ML, Dutang C. fitdistrplus: An R package for fitting distributions. *J. Stat. Softw.* 2015; **64**(4): 1–34.
39. Dillner AM, Takahama S. Predicting ambient aerosol thermal–optical reflectance (TOR) measurements from infrared spectra: organic carbon. *Atmos. Meas. Tech.* 2015; **8**(3): 1097–1109.
40. Chow JC, Watson JG, Pritchett LC, Pierson WR, Frazier CA, Purcell RG. The DRI thermal optical reflectance carbon analysis system - description, evaluation and applications in United-States air-quality studies. *Atmos. Environ. Part A-general Topics* 1993; **27**(8): 1185–1201.
41. Russell LM. Aerosol organic-mass-to-organic-carbon ratio measurements. *Environ. Sci. Technol.* 2003; **37**(13): 2982–2987.
42. Takahama S, Johnson A, Russell LM. Quantification of carboxylic and carbonyl functional groups in organic aerosol infrared absorbance spectra. *Aerosol Sci. Technol.* 2013; **47**(3): 310–325.
43. Pavia D, Lampman G, Kriz G. *Introduction To Spectroscopy*. Brooks/Cole Pub Co: Belmont, CA, 2008.
44. Hawkins LN, Russell LM, Covert DS, Quinn PK, Bates TS. Carboxylic acids, sulfates, and organosulfates in processed continental organic aerosol over the southeast Pacific Ocean during VOCALS-REx 2008. *J. Geophys. Res. Atm.* 2010; **115:** D13201, DOI:10.1029/2009JD013276.
45. Subramanian R, Khlystov AY, Cabada JC, Robinson AL. Positive and negative artifacts in particulate organic carbon measurements with denuded and undenuded sampler configurations. *Aerosol Sci. Technol.* 2004; **38**: 27–48.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.

Supporting Information for “Model selection for partial least squares calibration and implications for analysis of atmospheric organic aerosol samples with mid-infrared spectroscopy”

Satoshi Takahama and Ann M. Dillner

S1 Additional description of *M2*

If the variance and RVM increases monotonically with model complexity (characterized by number of LVs for PLS) as shown in Figure 2 for our study, $\mathit{RVM}_1 = \min_k\{\mathit{RVM}_k\}$, $\mathit{RVM}_{k^*} = \max_k\{\mathit{RVM}_k\}$, and $\mathit{M2}_{k^*} = 1 + \lambda$. The values for *M2* in such a case is illustrated in Figure S1. Even if the increase is not strictly monotonic [e.g., 1], the upper bound in the growth in RMSE indicated by Equation 7 will still hold.

We can present an alternative formulation in which $\phi = \lambda/\lambda^*$ bounds the anticipated increase in $\mathit{RMSE}_{k^\dagger}$ with respect to the estimated magnitude of achievable reduction in prediction errors (Section 2.1). However, the solution selected using this parameterization, denoted as *M2'*, is identical to that selected by *M2*; therefore, we can use *M2* and calculate the equivalent ϕ from λ if desired. To confirm that this is true, let us define *M2'* as

$$\mathit{M2}'_k = \left(\frac{\mathit{RMSE}_k - \min_k\{\mathit{RMSE}_k\}}{\max_k\{\mathit{RMSE}_k\} - \min_k\{\mathit{RMSE}_k\}}\right) + \phi\left(\frac{\mathit{RVM}_k - \min_k\{\mathit{RVM}_k\}}{\max_k\{\mathit{RVM}_k\} - \min_k\{\mathit{RVM}_k\}}\right)$$

where $\{x_k\}$ denotes the set of values for all models $\{x_k : k = 1, 2, \ldots, \kappa = k^*\}$. We can see that $\lambda^* \mathit{M2}'_k = \mathit{M2}_k - 1$; therefore $k^\dagger = \arg\min_k\{\mathit{M2}'_k\} = \arg\min_k\{\lambda^*\mathit{M2}'_k\} = \arg\min_k\{\mathit{M2}_k - 1\} = \arg\min_k\{\mathit{M2}_k\}$. With this formulation, we can additionally see that when $\phi = 1$, $k^\dagger = \arg\min_k\{\mathit{M2}_k\} = \arg\min_k\{\mathit{M1}_k\}$ under the conditions that $\lambda = \lambda^*$ and $\kappa = k^*$.

S2 Comparison of *M1* and *M2*

By analogy to Figure 3 generated for *M2*, in Figure S2 we reflect on the variation of $k^\dagger$ and corresponding fit metrics as a function of $\kappa = \{k^*, k^* + 1, \ldots, 120\}$ as formulated by *M1*. $k^\dagger = \arg\min_k\{\mathit{M1}_k\}$ is smallest when $\kappa = k^*$, and increases with increasing κ. For cCOH and CO, we find that the minimum RMSECV solution is reached with *M1* when $\kappa \approx 2k^*$. For aCOH, the selected solution is invariant after $\kappa \approx 2k^*$ but does not correspond to the RMSECV solution, though the value of RMSECV approaches that of its minimum. For aCH, the number of factors selected is invariant over the domain explored from $\kappa = k^*$ to 120, even while $\|\hat{\boldsymbol{b}}\|$ceases to increase substantially after $k = 100$ (Figure 2). The increase in RMSECV and RMSEP for the models selected as κ is varied can range between 0 and 365%, presumably from an increasing degree of under-fitting.

Comparison of the selected number of LVs according to parameters of both *M1* and *M2* is shown in Figure S3. This figure also shows that $k = k^*$ at $\kappa > 60$ for cCOH and CO (as they correspond to the solution where $\lambda = 0$), whereas $k > k^*$ for $\kappa \leq 120$ for cCOH and aCH. The selected number of LVs stabilizes as κ is increased between 60 and 120 (and possibly beyond), and corresponds to values similar to those selected by ensemble scoring (Table 1). It is conceivable that model performance curves generated by *M1* an ensemble of κ values can also be used with SRD for model selection.

References

1. Gowen, A. A., Downey, G., Esquerre, C. & O'Donnell, C. P. Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *Journal of Chemometrics* **25,** 375–381 (2011).
2. Ruthenburg, T. C., Perlin, P. C., Liu, V., McDade, C. E. & Dillner, A. M. Determination of organic matter and organic matter to organic carbon ratios by infrared spectroscopy with application to selected sites in the IMPROVE network. *Atmospheric Environment* **86,** 47–57 (2014).
3. Dillner, A. M. & Takahama, S. Predicting ambient aerosol thermal-optical reflectance (TOR) measurements from infrared spectra: organic carbon. *Atmospheric Measurement Techniques* **8,** 1097–1109 (2015).

Tables

Table S1: Comparison of minimum RMSE solution between Ruthenburg *et al.* [2][a] and this work[b].

	Number of LVs			
	aCOH	cCOH	aCH	CO
Ruthenburg *et al.* [2]	22	27	19	27
This work	16	18	54	18

[a] Selected from $\kappa = 30$. In the publication, values of 20, 20, 12, 16 were incorrectly used for aCOH, cCOH, aCH, and CO, respectively, due to user error. The correct number of LVs were used for calculation of organic matter to organic carbon ratios (OM/OC) used by Dillner & Takahama [3].

[b] Selected from $\kappa = 120$.

Figures

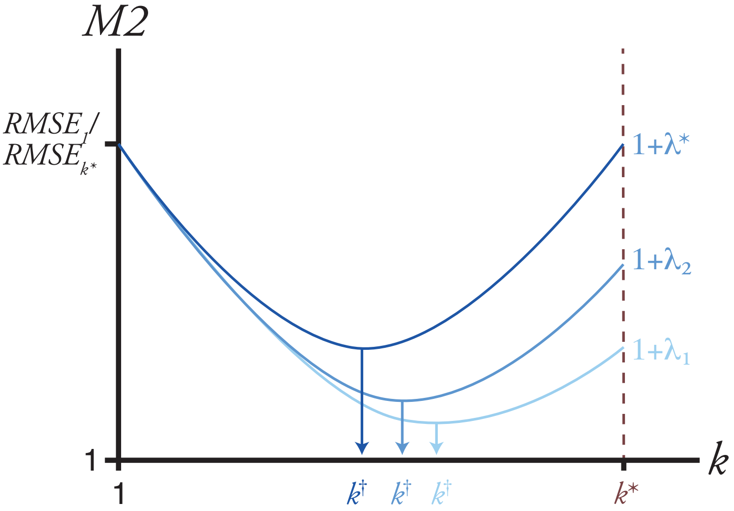

Figure S1: Illustration of the proposed $M2$ metric over its domain for several hypothetical values of λ ($\lambda_1 < \lambda_2 < \lambda^*$). k is the number of LVs, $k^* = \arg\min_k\{RMSE_k\}$, and $k^\dagger = \arg\min_k\{M2_k\}$. When $\lambda = \lambda^* = \max_k\{RMSE_k\}/\min_k\{RMSE_k\} - 1$ (top curve), $M2_1 = M2_{k^*} = \max_k\{RMSE_k\}/\min_k\{RMSE_k\}$ if $\max_k\{M2_k\} = M2_1$.

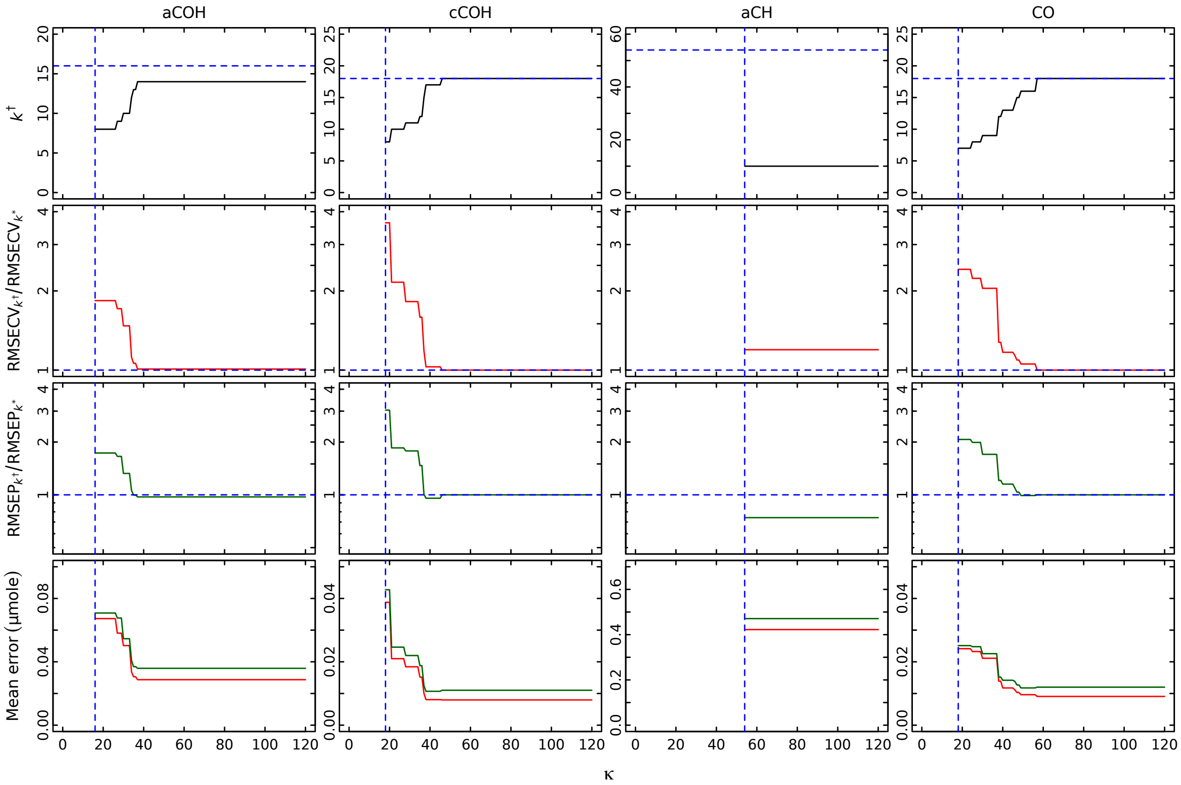

Figure S2: The number LVs selected according to $k = k^{\dagger}$ (top row) and corresponding fit metrics for laboratory standards as the parameter κ is varied from k^{*} to 120 (rows 2–4). Dotted blue horizontal and vertical lines correspond to the $k = k^{*}$ solution. Solid red lines are used to indicate evaluations for calibration samples, and solid green lines indicate evaluations for test set samples. Note that the y-axes for panels in rows 2 and 3 are in logarithmic scale, and share the same limits across all columns.

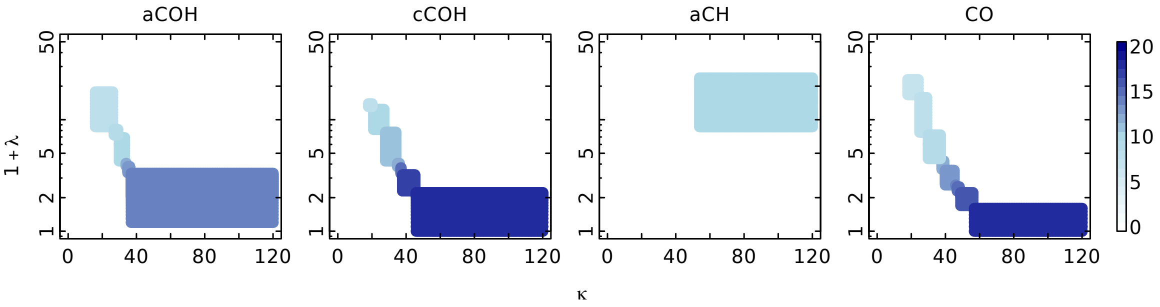

Figure S3: Correspondence of κ and $1 + \lambda$ which results in the same number of LVs selected by *M1* and *M2*, respectively. Colors (blue gradient) indicates the number of LVs selected at each coordinate.

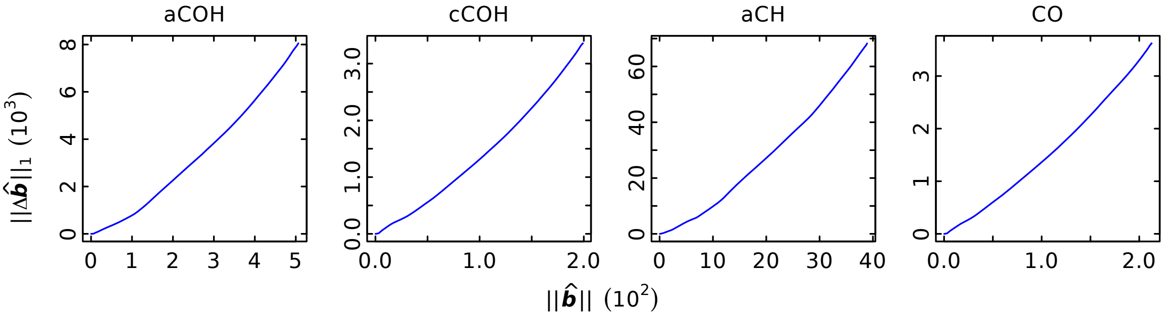

Figure S4: Comparison of RVMs: $\|\mathit{\Delta}\hat{\boldsymbol{b}}\|_1$ and $\|\hat{\boldsymbol{b}}\|$ for $k = \{1, 2, \ldots, \kappa = 120\}$.

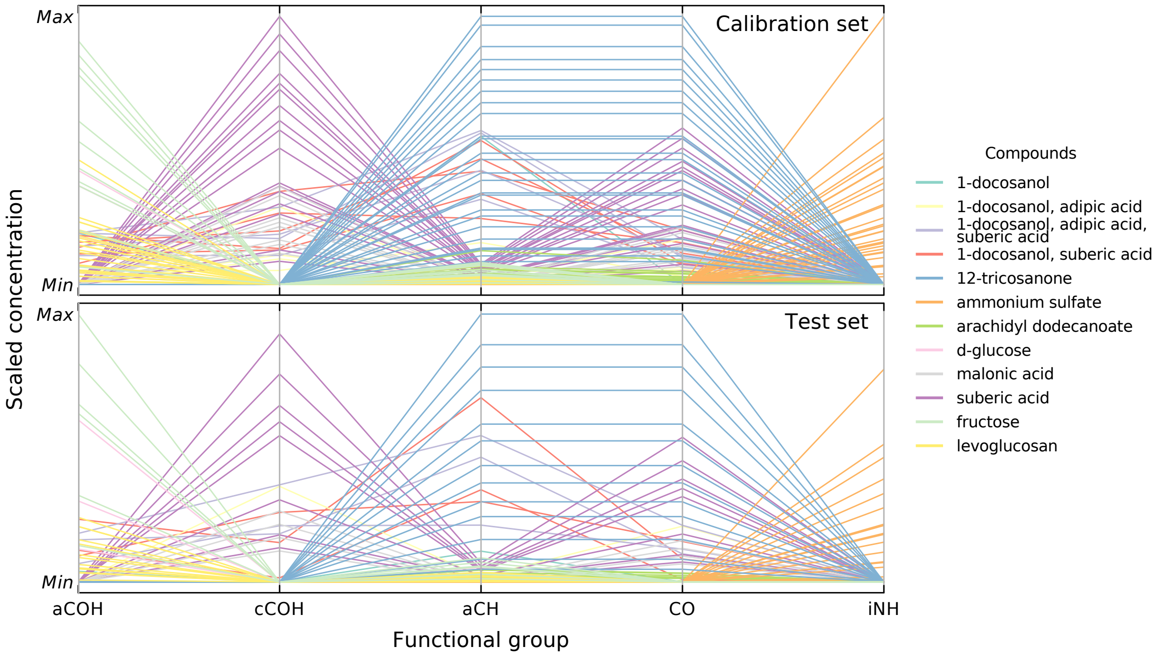

Figure S5: Parallel coordinate plots of functional group concentrations in laboratory standards (calibration and test sets). For each functional group, concentration in μmoles is scaled by subtracting the minimum value and divided by the range computed for all samples. Lines connect abundances observed in each sample, and are colored according to the compounds present in the standard. iNH is inorganic NH contained in ammonium sulfate included in the calibration set. Ammonium sulfate is included in the standards, as it is an interferant for quantification of aCOH, cCOH, and aCH due to overlapping absorption bands of the iNH bond.