Research Article

Journal of CHEMOMETRICS

Received: 1 October 2010, Revised: 21 March 2011, Accepted: 1 April 2011, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/cem.1398

Orthogonal projection to latent structures solution properties for chemometrics and systems biology data

David J. Biagioni[a]*, David P. Astling[b], Peter Graf[b] and Mark F. Davis[b]

Partial least squares (PLS) is a widely used algorithm in the field of chemometrics. In calibration studies, a PLS variant called orthogonal projection to latent structures (O-PLS) has been shown to successfully reduce the number of model components while maintaining good prediction accuracy, although no theoretical analysis exists demonstrating its applicability in this context. Using a discrete formulation of the linear mixture model known as Beer's law, we explicitly analyze O-PLS solution properties for calibration data. We find that, in the absence of noise and for large *n*, O-PLS solutions are simpler but just as accurate as PLS solutions for systems in which analyte and background concentrations are uncorrelated. However, the same is not true for the most general chemometric data in which correlations between the analyte and background concentrations are nonzero and pure profiles overlap. On the contrary, forcing the removal of orthogonal components may actually degrade interpretability of the model. This situation can also arise when the data are noisy and *n* is small, because O-PLS may identify and model the noise as orthogonal when it is statistically uncorrelated with the analytes. For the types of data arising from systems biology studies, in which the number of response variables may be much greater than the number of observations, we show that O-PLS is unlikely to discover orthogonal variation whether or not it exists. In this case, O-PLS and PLS solutions are the same. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: O-PLS; partial least squares (PLS); Mid-infrared (MIR) calibration; Beer's law; systems biology

1. INTRODUCTION

Partial least squares (PLS), sometimes referred to as projections to latent structures, is one of the most commonly used multivariate regression methods in chemometrics [1–3]. PLS is designed for applications in which the data have more independent variables than observations (the "small *n*, large *p*" regime) and where collinearity is present among variables. In contrast to ordinary least squares (OLS) regression, which is ill posed in the "small *n*, large *p*" regime, PLS assumes that meaningful structure in the data is low dimensional and attempts to identify as few linear combinations of the independent variables (*latent variables*) as possible without sacrificing prediction quality.

Nonetheless, interpretation of PLS components can be challenging in cases where significant systematic variation is present in the measured data. Such variation can arise from a variety of sources, including temperature fluctuations, instrument and sample handling errors, and contamination of interfering substances. Even in well-designed calibration studies, where the response variables are carefully controlled to isolate corresponding changes in the spectrum, systematic variation may still present challenges in interpreting the PLS components and may adversely affect prediction. Although one can often improve PLS prediction accuracy by introducing more components to the model, doing so has an adverse effect on model interpretability and can lead to poorer prediction quality due to overfitting. For example, one recent study used PLS for calibration and estimation of algal lipid content from mid-infrared (Mid-IR) spectral data [4]. Despite the small number of changing experimental factors, a relatively large number of PLS components were needed to adequately characterize the variation in the spectral data while preserving accurate predictions of lipid concentration.

In 2002, Trygg and Wold [5] introduced a variant of PLS, called O-PLS, that was designed to remove systematic variation uncorrelated with the response. The idea was that doing so reduces the number of predictive components, thereby increasing interpretability of the model factors. Indeed, in calibration settings, O-PLS models are known to have similar prediction quality to PLS but often with far fewer, more meaningful components. For nonlinear problems, the algorithm has been adapted for use with the "kernel trick [6]." Another variant, O2-PLS, works in much the same way as O-PLS but is applied bi-directionally [7]. The data blocks are treated as predictor and response in turn to identify both **X**-orthogonal variation in **Y** and **Y**-orthogonal variation in **X**. The O2-PLS model is constructed by applying O-PLS to the data in both directions, and then building additional linear models relating the **X**-scores and **Y**-scores.

The success of O-PLS in spectroscopy has led to its use in other fields. Recently, O2-PLS was applied in a systems biology

* *Correspondence to: David J. Biagioni, Department of Applied Mathematics, University of Colorado at Boulder, 526 UCB, Boulder, CO 80309-0526, USA. E-mail: biagioni@colorado.edu*

a *D. J. Biagioni*
Department of Applied Mathematics, University of Colorado at Boulder, 526 UCB, Boulder, CO 80309-0526, USA

b *D. P. Astling, P. Graf, M. F. Davis*
National Renewable Energy Laboratory (NREL), 1617 Cole Blvd, Golden, CO 80401-3305, USA

context to integrate metabolomic and transcriptomic data [8]. The same authors have also applied it to three-way data to find relationships among proteomic, transcriptomic, and metabolomic profiles [9]. Unfortunately, these studies never compared the O2-PLS models with analogous PLS models for the same data.

To date, theoretical descriptions of O-PLS solutions have mostly focused on their relationship to PLS, without reference to any underlying model describing the data. For instance, Kemsley and Tapp [10] have suggested a simple method for filtering **X** using only the PLS solution. This result is consistent with prior observations by Ergon [11], who has also proposed projection-based techniques to reduce the number of PLS model components [12]. Verron *et al.* [13] have also made interesting observations regarding the connections between O-PLS components and their PLS counterparts. Our contribution, and the central theme of this paper, entails an analysis of O-PLS applicability to chemometric and systems biology data, which, to our knowledge, has not been made in the literature. We place special emphasis on the qualitative properties of O-PLS solutions, often assuming that the data are noise free and that n is large enough for tight covariance estimates to be valid. In doing so, we explicitly outline solution properties while highlighting some of the algorithm's strengths and limitations. In particular, following the approach set by Nadler and Coifman [14,15], we begin by analyzing O-PLS performance on spectral data obeying Beer's law and show explicitly why the algorithm often produces informative results in this context. We extend this discussion to the systems biology context in which one is likely to encounter data matrices with very different properties than in calibration studies and highlight the challenges this presents to the O-PLS algorithm.

1.1. Comparison of partial least squares and orthogonal projection to latent structures

In order for the discussion to be reasonably self-contained, we provide a brief background on the mathematical formulation of PLS and O-PLS. Let **X** and **Y** be $n\times p$ and $n\times m$ matrices, respectively, where p, m are the numbers of variables in each data set and n is the number of observations. We assume that **Y** is well approximated by

$$\hat{\mathbf{Y}} = \mathbf{X}\beta \tag{1.1}$$

where β is the $p\times m$ matrix of regression coefficients. Although both $\hat{\mathbf{Y}}$ and β are understood to be sample estimates, we use a hat only on $\hat{\mathbf{Y}}$ to distinguish it from the data **Y**. No such clarification is needed for the matrix β, which we always understand to be an estimator. When $n > p$ (i.e., the problem is *overdetermined*) and **X** is of full rank, OLS solution, β_{OLS}, is found via the normal equations,

$$\beta_{\mathrm{OLS}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

When $n < p$ (i.e., the problem is *underdetermined*), the matrix $\mathbf{X}^T\mathbf{X}$ is guaranteed to be singular and no unique β_{OLS} satisfies (1.1).

Partial least squares assumes that meaningful structure in the data is inherently low dimensional, so that only a small subset of the predictor variables is necessary to predict **Y**. These subsets are traditionally called *latent variables* because they are, in principle, unmeasurable by themselves [2]. A succinct form of the PLS model is given by

$$\begin{aligned}\mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E}\\ \hat{\mathbf{Y}} &= \mathbf{T}\mathbf{C}^T\end{aligned} \tag{1.2}$$

where the matrices $\mathbf{T}, \mathbf{P}$ and **C** are of relatively low rank. By analogy with principal components analysis (PCA), the columns of matrix **T** are called the *scores*, those of **P** are the *loadings* (or **X**-*loadings*), those of **C** are the **Y**-*loadings*, and **E** is the *residual* matrix. The scores represent the coordinates of the observations with respect to the loadings, **W**, of PCA(**K**), where $\mathbf{K} = (1/n)\,\mathbf{Y}^T\mathbf{X}$ is the sample covariance matrix for mean-centered data. The **W**-loadings and **Y**-loadings are then computed so that each component is a best least squares fit. Eqn 1.2 is easily rearranged to find that the PLS regression coefficient depends on both the **X**-loadings and **Y**-loadings via

$$\beta_{\mathrm{PLS}} = \mathbf{W}\mathbf{C}^T \tag{1.3}$$

O-PLS may be thought of as PLS combined with a preprocessing step that filters systematic variation from **X** that is orthogonal, or statistically uncorrelated, to the **Y** variables. The O-PLS model has the form

$$\begin{aligned}\mathbf{X} &= \mathbf{T}\mathbf{W}^T + \mathbf{T}_{\mathrm{orth}}\mathbf{P}_{\mathrm{orth}}^T + \mathbf{E}\\ \hat{\mathbf{Y}} &= \mathbf{T}\mathbf{C}^T\end{aligned}$$

where the subscript $(\cdot)_{\mathrm{orth}}$ denotes orthogonal components and, in general, **T** and **C** are not the same as those of PLS. The regression matrix for O-PLS, $\beta_{\text{O-PLS}}$, is found as in Eqn 1.3 but, given the underlying dependence of **C** on the scores, will generally differ from the PLS regression matrix. New predictions are obtained in the standard way via Eqn 1.1 after first filtering the new samples of orthogonal variation.

For reference, we summarize the steps of O-PLS following Trygg and Wold [7].

(1) Calculate A PCA components of the $m\times p$ covariance matrix

$$PCA(\mathbf{K}) = PCA\left(\frac{1}{n}\mathbf{Y}^T\mathbf{X}\right) = \mathbf{T}_{\mathrm{PCA}}\mathbf{W}^T \tag{1.4}$$

where $\mathbf{T}_{\mathrm{PCA}}$ is an $m\times A$ score matrix and **W** is a $p\times A$ loading matrix with orthonormal columns. Recall that $A \le m$ and that we are primarily interested in cases where $m << p$.[1]

(2) Calculate the $n\times A$ predictive score matrix $\mathbf{T} = \mathbf{X}\mathbf{W}$. These are coordinates of the rows of **X** with respect to the orthonormal basis in **W**.
(3) Calculate the $n\times p$ residual matrix, $\mathbf{E}_{xy} = \mathbf{X} - \mathbf{T}\mathbf{W}^T = \mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^T$. This step subtracts away the projection of the rows of **X** onto the basis in **W**.
(4) Calculate $\mathbf{w}_{\mathrm{orth}}$, the most significant loading of PCA($\mathbf{T}^T\mathbf{E}_{xy}$).
(5) Sequentially remove structured noise, $\mathbf{t}_{\mathrm{orth}}\mathbf{p}_{\mathrm{orth}}^T$, from **X**:

$$\begin{aligned}\mathbf{t}_{\mathrm{orth}} &= \mathbf{X}\mathbf{w}_{\mathrm{orth}}\\ \mathbf{p}_{\mathrm{orth}}^T &= \mathbf{t}_{\mathrm{orth}}^T\mathbf{X}/\|\mathbf{t}_{\mathrm{orth}}\|^2\\ \mathbf{X} &\leftarrow \mathbf{X} - \mathbf{t}_{\mathrm{orth}}\mathbf{p}_{\mathrm{orth}}^T\end{aligned}$$

[1] In the event that $m = 1$, then $A = 1$ and there is no need to perform PCA. This is the case in the example in Section 2.4.

(6) For additional orthogonal components, repeat steps 4 and 5. Otherwise, repeat step 2 using the filtered **X** to update the score matrix **T**, and go to step 7.
(7) Construct a PLS model from **Y** and the filtered **X**.
(8) To predict using a new sample, $\mathbf{x}^{(n+1)}$, first filter any orthogonal variation. Then $\hat{y}^{(n+1)} = \mathbf{x}^{(n+1)}\beta_{\text{O-PLS}}$.

The authors show that the algorithm maximizes the L_2-norm of the projection of each $\mathbf{t}_{\text{orth}}$ onto **T**, thereby removing as much structured noise from the **Y**-predictive components as possible. Specifically, $\mathbf{t}_{\text{orth}}$ satisfies the optimization criteria

$$\max_{\mathbf{t}} \|\mathbf{T}^T\mathbf{t}\|_2 \text{subject to } \mathbf{Y}^T\mathbf{t} = 0. \tag{1.5}$$

O-PLS models often have fewer components than PLS for a given accuracy. This is due to the removal of **Y**-orthogonal components, which, although potentially useless for prediction, may still describe a large fraction of the total variation in **X**. In this sense, we understand such O-PLS solutions to be *more interpretable* than the PLS ones: if fewer components are needed for equivalent prediction quality, the corresponding latent variables should be easier to interpret.

For a description of how O-PLS relates to O2-PLS, we refer the reader to the original publication on the subject [7]. Because of the symmetric nature of the O2-PLS formulation (first **X** is the predictor and **Y** the response, then vice versa) and to keep the derivations tractable, we focus only on O-PLS. We do so with the understanding that all arguments also apply in reverse, with **X** considered the response matrix and **Y** the predictor matrix.

2. ORTHOGONAL PROJECTION TO LATENT STRUCTURES SOLUTION PROPERTIES FOR DATA OBEYING BEER'S LAW

2.1. Beer's law and calibration data

The original O-PLS literature considered the application of the algorithm to spectroscopic data [5]. In this section, we show explicitly not only why O-PLS results may be easier to interpret than PLS results for Mid-IR calibration data, but also how certain undesirable solution features can arise depending on various properties of the data. By imposing different restrictions on the input data, we can analyze directly the strengths and limitations of the algorithm.

The theoretical foundation for Mid-IR calibration is Beer's law, which states that the spectral data are a linear mixture of the pure constituent profiles weighted by their concentrations. In an appropriately chosen system of units, the law may be written as a sum of outer products of all constituent concentration and pure profile vectors,

$$\mathbf{X} = \sum_{i=1}^{s} \mathbf{y}_i\mathbf{z}_i^T \tag{2.1}$$

Here, **X** is the $n \times p$ matrix of spectral data, s is the total number of constituents in the sample, $\mathbf{y}_i$ is an n-vector of concentrations, and $\mathbf{z}_i$ is the p-dimensional pure constituent profile (extinction coefficients) of species i. It is most often the case in calibration studies that only a relatively small number j of the s concentrations are of interest, making it convenient to separate the expression in Eqn 2.1 into two sums, the first corresponding to the analytes and the second to the background constituents whose properties are, in general, unknown. We assume the indices have been chosen so that the first j components $i = 1, \ldots, j$ are the analytes and the next $k = s - j$ components $i = j+1, j+2, \ldots, s$ are the background constituents. We continue to use the symbols **y** and **z** to denote analyte concentrations and spectral profiles but choose **u** and **v** to denote the analogous background properties. Beer's law may then be expressed as

$$\mathbf{X} = \sum_{i=1}^{j} \mathbf{y}_i\mathbf{z}_i^T + \sum_{i=1}^{k} \mathbf{u}_i\mathbf{v}_i^T \tag{2.2}$$

with $s = j + k$. To simplify the notation one step further, we matricize the summations in Eqn 2.2 to acquire

$$\mathbf{X} = \mathbf{Y}\mathbf{Z}^T + \mathbf{U}\mathbf{V}^T \tag{2.3}$$

where **Y** ($n \times j$) and **Z** ($p \times j$) contain the $\{\mathbf{y}_i\}_{i=1}^{j}$ and the $\{\mathbf{z}_i\}_{i=1}^{j}$ in their columns, respectively, and similarly for **U** ($n \times k$), **V** ($p \times k$), $\{\mathbf{u}_i\}_{i=1}^{k}$, and $\{\mathbf{v}_i\}_{i=1}^{k}$.

Some authors have used a probabilistic model of the input data, where the measured concentrations are samples from a potentially noisy underlying distribution [14,15]. Although we recognize the theoretical value of a probabilistic approach, we feel that most researchers actively using PLS and O-PLS/O2-PLS will be more familiar with a discrete formulation. That said, statistical estimates for finite training data may contain large variance, especially when n is small. Because we are primarily interested in the qualitative properties of O-PLS solutions, we will often assume that n is large enough for certain covariance estimates (e.g., $\text{Cov}(\mathbf{Y}, \mathbf{U}) = 0$) to be valid. The only exception to this is in Section 2.3, in which we examine how the algorithm handles noisy data as a function of n and the noise strength.

In the following, we assume mean centering of the data and denote the covariance of two matrices, **A** and **B**, by $\text{Cov}(\mathbf{A}, \mathbf{B}) = \frac{1}{n}\mathbf{A}^T\mathbf{B}$. With some abuse of notation, we define analogous quantities between matrices and vectors, and vectors and vectors. We also note that variance and covariance are related in the usual way, so that for any vector **a**, $\text{Var}(\mathbf{a}) = \text{Cov}(\mathbf{a}, \mathbf{a})$.

2.2. Uncorrelated concentrations and non-overlapping pure profiles

We first consider the simplest possible assumptions about data obeying Beer's law (Eqn 2.3). We show that PLS and O-PLS solutions are identical, in particular because O-PLS does not find **Y**-orthogonal variation. The assumptions are then relaxed to obtain more general results. Unless otherwise stated, we take the data to be noise free and n to be "large enough" to justify ignoring high-order terms in the covariance estimates (although we do briefly describe how O-PLS handles noise for small n in Section 2.3).

Suppose the system has the following properties:

Assumption 1: uncorrelated concentrations, $\text{Cov}(\mathbf{Y}, \mathbf{U}) = 0$

Assumption 2: non-overlapping pure profiles, $\mathbf{Z}^T\mathbf{V} = 0$.

Physically, these assumptions imply that the analyte concentrations change completely independently of the background across all samples, and that their pure profiles share no spectral peaks with the background. Although this situation may be unlikely in real experiments, it is a useful starting point for the analysis because of its simplicity.

First, we note that both PLS and O-PLS covariance loadings, **W**, are linear combinations of the pure analyte profiles. To see this, recall that the first step of both algorithms computes the **XY**-covariance matrix, which, by Assumption 1, reduces to

$$\mathbf{K} = \frac{1}{n}\mathbf{Y}^T\mathbf{X} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}\mathbf{Z}^T = \mathrm{Cov}\,(\mathbf{Y})\mathbf{Z}^T \tag{2.4}$$

From this, it follows that the columns of the $p \times A$ loading weight matrix, **W**, of PCA(**K**) must be linear combinations of the pure profiles and hence lie in the subspace $\mathrm{Span}\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_j\}$, having dimension d_j. Then, by Assumption 2, the background pure profiles must be orthogonal to all of the loading weights, $\mathbf{V}^T\mathbf{W} = 0$.

Second, both PLS and O-PLS find identical score matrices whose columns are linear combinations of the analyte concentrations,

$$\mathbf{T} = \mathbf{X}\mathbf{W} = \mathbf{Y}\mathbf{Z}^T\mathbf{W} = \mathbf{Y}\mathbf{A}_T \tag{2.5}$$

weighted by the spectral projection matrix $\mathbf{A}_T = \mathbf{Z}^T\mathbf{W}$. This is desirable from a prediction standpoint because the background constituents are assumed to be uncorrelated with the analytes. Note that when there is only one analyte ($j = 1$), Eqn 2.5 implies that the first score is exactly proportional to the analyte concentration.

Under the given conditions, the first orthogonal component of O-PLS is zero. To see this directly, first compute the **X**-residual,

$$\mathbf{E}_{xy} = \mathbf{X} - \mathbf{T}\mathbf{W}^T = \left(\mathbf{Y}\mathbf{Z}^T + \mathbf{U}\mathbf{V}^T\right) - \mathbf{Y}\mathbf{Z}^T\mathbf{W}\mathbf{W}^T = \mathbf{U}\mathbf{V}^T$$

where we have used the fact that $\mathbf{Z}^T = \mathbf{Z}^T\mathbf{W}\mathbf{W}^T$ because **W** contains a basis for the columns of **Z**. This and Assumption 1 yield the result directly,

$$\frac{1}{n}\mathbf{E}_{xy}^T\mathbf{T} = \mathbf{V}\,\mathrm{Cov}\,(\mathbf{U}, \mathbf{Y})\mathbf{A}_T = 0 \tag{2.6}$$

No **Y**-orthogonal variation is found, and both PLS and O-PLS produce exactly the same predictive model. Omitting the simple derivations for brevity, we note that for a single response under these conditions, the O-PLS solution has the following properties:

$$\mathbf{t} \propto \mathbf{y}, \quad \mathbf{p}, \beta_{\text{O-PLS}} \propto \mathbf{z}, \quad \mathbf{w}_{\text{orth}} = 0$$

2.3. Relation to Gaussian noise

We note that when n is small, the covariance estimate in Assumptions 1 and 2 will be $O(1/\sqrt{n})$. A careful look at the steps of the algorithm reveals that this leads to correction terms of $O(1/\sqrt{n})$ in Eqn 2.6. Any implementation of O-PLS will have an orthogonality tolerance to determine whether the matrix in Eqn 2.6 contains significant components. So long as the error in the covariance estimate stays below this tolerance, the results of the previous paragraphs remain unchanged.

Regarding the algorithm's handling of noise, the authors of O-PLS claim that "results from initial studies with O-PLS do not show any degradation of results compared to non-treated data" [5]. Indeed, in the limit of large n when the covariance estimates are precise, noise is uncorrelated with the analyte concentrations and O-PLS classifies it as the residual. For small n, however, it is possible that O-PLS will identify noise as being orthogonal to **Y**.

We make these statements more precise and illustrate with a simple example. Suppose that the data are modeled by Beer's law (Eqn 2.3) with noise,

$$\mathbf{X} = \mathbf{Y}\mathbf{Z}^T + \mathbf{U}\mathbf{V}^T + \sigma\xi$$

where σ is a non-negative parameter characterizing noise strength and ξ is a matrix whose elements are independent identically distributed samples from $\mathcal{N}(0, 1)$. For large enough n, Assumptions 1 and 2 remain valid, and the given results still hold. In particular, no orthogonal variation is identified by O-PLS, and the O-PLS and PLS models are the same.

For small n, however, we must instead assume that $\mathrm{Cov}\,(\mathbf{Y}, \xi) = O(1/\sqrt{n})$. Although a detailed analysis of the solution properties' dependence on σ, n, and p is beyond the scope of this paper, we can demonstrate that the following properties are true for a single-component system with Gaussian noise:

(1) $\max|\mathbf{w}_{\text{orth}}| = O(\sigma^2/n)$. This quantity is directly related to the tolerance used by O-PLS to identify orthogonal variation.
(2) Filtering of noise by O-PLS can lead to lower residuals, but at the cost of overfitting.

We omit a detailed proof of these statements and instead present a numerical example to support the claims.

Figure 1 shows O-PLS solution properties for a single-component system with noise. The noise parameter σ is allowed to vary from 0 to 0.04, where $\sigma = 0$ corresponds to a perfectly clean signal and $\sigma = 0.04$ to a highly corrupted one (left panel). The center panel provides numerical evidence for the claim that $max\,|\mathbf{w}_{\text{orth}}| = O(\sigma^2/n)$. Note that on a log–log scale and for all fixed n, the dependence of $max\,|\mathbf{w}_{\text{orth}}|$ on sigma is approximately linear with slope 2 (it is quadratic in σ). On the other hand, for all fixed σ, the spacing between the lines is approximately constant. Because $n = 2^l$ for $l = 1, \ldots, 6$, this fact implies that $max\,|\mathbf{w}_{\text{orth}}|$ is also proportional to $1/n$.

Recall that magnitude of $\mathbf{w}_{\text{orth}}$ determines whether O-PLS identifies orthogonal variation (step 4 of the O-PLS pseudocode). Hence, as this quantity increases, we would expect O-PLS to identify more orthogonal components. This is illustrated in the right panel of Figure 1, which shows the average number of orthogonal components identified as a function of σ for $n = 64$. In region *A*, corresponding to small σ, no orthogonal variation is found. In region *B*, either zero or one orthogonal component is identified, depending on the specific instance of the noise. As σ increases further (region *C*), O-PLS consistently identifies one orthogonal component, and so on for regions *D* and *E*. Although not shown for clarity, similar trends are apparent for other values of n, with the cutoff for each region shifted further to the left (smaller σ) for smaller n. Although the residual for O-PLS decreases as the number of orthogonal components increases, this may be viewed as overfitting because the algorithm is building a model of the noise that will be used to filter future samples.

2.4. Uncorrelated concentrations and overlapping pure profiles

The assumption of uncorrelated concentrations may be valid in calibration experiments where the known constituents are added directly to an otherwise homogeneous sample set (i.e., in "spike-in" calibration studies [4]). Although the concentrations of the analytes vary, those of the unknown background constituents should remain roughly the same from sample to sample, resulting in zero or near-zero sample correlations between analyte and background concentrations. On the other hand, the second assumption of non-overlapping pure profiles is often invalid because the spectra of many constituents share peaks.

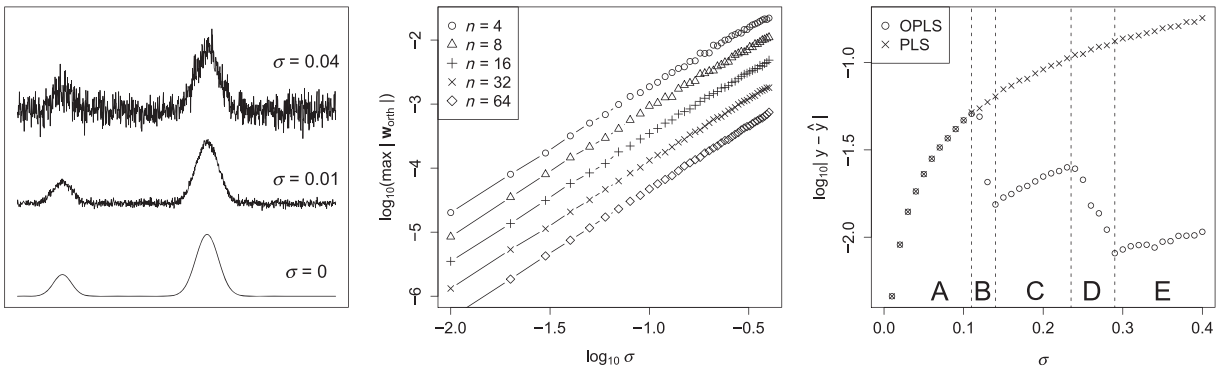

Figure 1. O-PLS model dependence on noise strength σ and sample size n for a simple system. The spectral data are of the form $X = yz^T + \sigma\Xi$, where y is mean centered and of unit variance and z is the pure profile with $max\ z = 2$. The noise is composed of a scalar σ and matrix ξ whose columns are random samples from a Gaussian distribution with mean zero and unit variance. Left: Dependence of sample spectra on noise parameter σ. Center: The max norm of w_{orth} is of order $O(\sigma^2/n)$. The norm of this vectors determines whether or not O-PLS identifies orthogonal components. Right: PLS and O-PLS log_{10} RMSEP as a function of σ over 50 trials for $n = 64$. Below the O-PLS threshold for w_{orth} (interval *A*), no orthogonal components are found, and O-PLS prediction is equivalent to PLS. As the magnitude of w_{orth} increases, O-PLS identifies one (interval *C*) and then two (interval *E*) orthogonal components. Intervals *B* and *D* are transition regions where the average number of orthogonal components falls between the adjacent values. Similar trends are apparent for other values of n, with the main difference being that the cutoff is shifted towards smaller σ for smaller n. Because the system has only a single non-noisy component, intervals B–E correspond to overfitting.

Therefore, we next examine PLS and O-PLS models under only Assumption 1 ($\mathrm{Cov}(\mathbf{Y}, \mathbf{U}) = 0$), allowing for the possibility of spectral overlap ($\mathbf{Z}^T\mathbf{V} \neq 0$). We find that O-PLS interpretability (corresponding to a smaller number of components) is more robust against increases in spectral overlap and concentration covariance than PLS. In particular, we will show that O-PLS solution components have the following properties:

(1) $\mathbf{w}_{\mathrm{orth}}^T\mathbf{Z} = 0$, that is, the orthogonal loading is orthogonal to the pure profiles.
(2) $\mathbf{t}_{\mathrm{orth}}$ is a linear combination of the background concentrations, $\{\mathbf{u}\}_{i=1}^{k}$.
(3) removing the orthogonal component, $\mathbf{p}_{\mathrm{orth}}\mathbf{t}_{\mathrm{orth}}^T$, removes only some linear combination of the background profiles.

Furthermore, for a single-response system, we derive the following properties:

$$\mathbf{t}_1 \propto \mathbf{y}, \mathbf{p} = \mathbf{z}, \beta_{\text{O-PLS}} = \check{\mathbf{z}}/\|\mathbf{z}\|$$

$$\mathbf{t}_{\mathrm{orth}} = \mathbf{u}, \mathbf{p}_{\mathrm{orth}} = \mathbf{v}, \mathbf{X}_1 = \mathbf{yz}^{\mathbf{T}}$$

where the inverted hat $\check{\cdot}$ denotes a unit vector.

To motivate the discussion, we first present simple numerical results for a three-component system in Figure 2. The system has one analyte and two background constituents with concentrations $\mathbf{y}$, $\mathbf{u}_1$, and $\mathbf{u}_2$ and spectral profiles $\mathbf{z}$, $\mathbf{v}_1$, and $\mathbf{v}_2$, respectively. Each column of the figure highlights qualitative properties of PLS and O-PLS solutions for a different degree of overlap of the pure profiles (overlap increases from left to right), on the assumption that all concentrations are uncorrelated (not shown). In this example, the single O-PLS component has a score vector collinear with $\mathbf{y}$ and a loading proportional to the pure profile. PLS, on the other hand, produces a three-component model with the same prediction accuracy as O-PLS, but whose loadings are more difficult to interpret. We outline the reasons for this below, occasionally referring to the figure to illustrate the key points.

The covariance matrix and its PCA loadings are the same as before for both PLS and O-PLS (2.4). However, the PLS score matrix is given by

$$\mathbf{T} = \mathbf{XW} = \mathbf{YB}_T + \mathbf{UC}_T \tag{2.7}$$

where we have defined $\mathbf{B}_T = \mathbf{Z}^T\mathbf{W}$ and $\mathbf{C}_T = \mathbf{V}^T\mathbf{W}$. Written in this way, Eqn 2.7 emphasizes how the PLS scores may contain contributions from potentially $\mathbf{Y}$-orthogonal background constituents. The extent to which this happens in general depends on the following:

- The variances of analyte and background concentrations. Because the variance of each individual concentration vector is proportional to its norm, columns of $\mathbf{Y}$ and $\mathbf{U}$ in Eqn 2.7 that have large variances will significantly affect the geometric direction and magnitude of the columns of $\mathbf{T}$.
- The parameter $a_T \equiv \|\mathbf{C}_T\|/(\|\mathbf{C}_T\| + \|\mathbf{B}_T\|)$, which we call the *profile overlap ratio* because it characterizes the extent to which the pure profiles of background constituents overlap with those of the calibration ones. Note that the ratio is zero when the background profiles are all orthogonal to the loading weights, $\mathbf{W}$, and approaches 1 as the background profiles lie increasingly in the subspace spanned by the basis in $\mathbf{W}$.

If either $\mathrm{Var}(\mathbf{U})$ or the overlap ratio is close to zero, the second term in Eqn 2.7 will not drastically contribute to the scores. In any other case, however, the second term may be non-negligible. Unfortunately, because the PLS score matrix contains contributions from $\mathbf{U}$, which is assumed to be orthogonal to $\mathbf{Y}$, the prediction accuracy of individual components is necessarily degraded. This type of variation, although not caused by measurement or sampling error, is noise with respect to prediction of the analytes and is exactly the type of variation O-PLS aims to characterize and remove (cf. third row of Figure 2).

Concentration covariance and spectral overlap also affect PLS loadings given by

$$\mathbf{P}^T = \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{X} \tag{2.8}$$

Although a closed-form expression for the inverse, $(\mathbf{T}^T\mathbf{T})^{-1}$ is intractable, we can rewrite the product $\mathbf{T}^T\mathbf{X}$ as

$$\mathbf{T}^T\mathbf{X} = \mathbf{B}_T^T\mathbf{Y}^T\mathbf{YZ}^T + \mathbf{C}_T^T\mathbf{U}^T\mathbf{UV}^T = \mathbf{B}_P\mathbf{Z}^T + \mathbf{C}_P\mathbf{V}^T$$

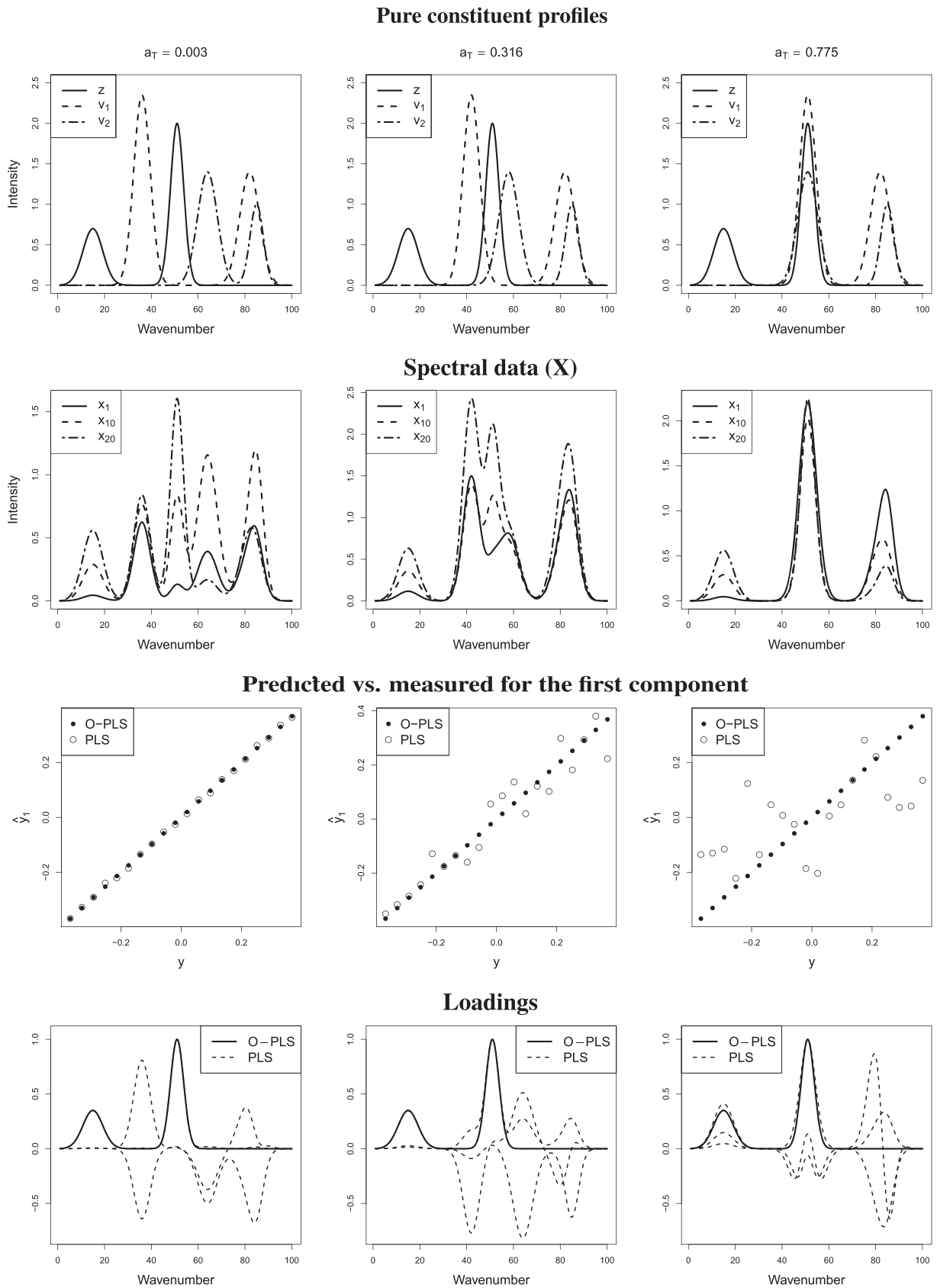

Figure 2. Partial least squares and O-PLS **y**-predictive models for three-component systems with varying degrees of overlap of the pure profiles, characterized by the profile overlap ratio, a_T, which increases from left to right. Analyte concentrations are assumed to be uncorrelated with the background concentrations. First row: Pure constituent profiles with various degrees of overlap. Second row: Three representative spectra. Third row: The first PLS and O-PLS scores (normalized for ease of visualization) are plotted against **y** for each value of a_T. Fourth row: Interpretability of PLS loadings is not straightforward (dashed lines) and changes significantly with a_T, whereas the single O-PLS loading remains proportional to the pure profile **z** independent of degree of overlap.

The $A \times j$ and $A \times k$ matrices $\mathbf{B}_P = n\mathbf{B}_T^T \mathrm{Cov}(\mathbf{Y})$ and $\mathbf{C}_P = n\mathbf{C}_T^T \mathrm{Cov}(\mathbf{U})$ are defined to simplify the expression. Written in this way, it is clear that the loadings depend on both **Z** and **V**. Hence, like the scores, PLS loadings may contain contributions from the unknown constituents, the strength of which depends in a complex way on spectral overlap and auto-covariance of the calibration and background constituent groups (cf. fourth row of Figure 2). Contrast this with the loadings described in Section 2.2 that depend only on the analyte profiles, **Z**.

Next, we show that O-PLS scores and loadings differ significantly from their PLS counterparts, resulting in fewer predictive components and better interpretability of the loadings. First, note that O-PLS will initially compute the same scores as PLS, $\mathbf{T} = \mathbf{Y}\mathbf{B}_T + \mathbf{U}\mathbf{C}_T$. It is straightforward to show that $\mathbf{w}_{\mathrm{orth}}$ is orthogonal to all of the pure profiles, $\mathbf{w}_{\mathrm{orth}}^T\mathbf{Z} = 0$. Hence, up to normalization, the first orthogonal score vector has the form

$$\mathbf{t}_{\mathrm{orth}} = \mathbf{X}\mathbf{w}_{\mathrm{orth}} = \mathbf{U}\mathbf{V}^T\mathbf{w}_{\mathrm{orth}} \tag{2.9}$$

The columns of $\mathbf{t}_{\mathrm{orth}}$ are linear combinations of the $\{\mathbf{u}_i\}_{i=1}^k$ only, meaning that each orthogonal component corresponds only to variation from the background. This is desirable because, by assumption, the background is uncorrelated with **Y** and should not contribute to the prediction model. In Figure 2, the single O-PLS predictive score is compared with the first PLS score in the third row. One can infer from these plots that the two orthogonal scores removed by O-PLS are some linear combination of the background concentrations because the only remaining score is perfectly collinear with **y**.

Analysis of the O-PLS loadings is not as simple as for PLS because they are computed after iteratively filtering **X**. However, it is useful to examine the orthogonal loadings that, up to normalization, have the form

$$\mathbf{p}_{\mathrm{orth}}^T = \mathbf{t}_{\mathrm{orth}}^T\mathbf{X} = n\mathbf{w}_{\mathrm{orth}}^T\mathbf{V}\,\mathrm{Cov}(\mathbf{U})\mathbf{V}^T \tag{2.10}$$

Clearly, each orthogonal component, $\mathbf{t}_{\mathrm{orth}}\mathbf{p}_{\mathrm{orth}}^T$, subtracts some linear combination of the pure background profiles from **X**. Again, the degree to which this happens depends in a complex way on the auto-covariance of the background concentrations and the degree of pure profile overlap. For our simple example, the O-PLS loading contains no contribution from the background profiles (Figure 2).

Closed-form analysis of O-PLS solutions for the most general case is beyond the scope of this paper. For simple systems with only a single response, however, it is easy to verify that O-PLS solutions have the mentioned properties.

For example, consider a two-component system,

$$\mathbf{X} = \mathbf{y}\mathbf{z}^T + \mathbf{u}\mathbf{v}^T$$

corresponding to Eqn 2.3 with $j = k = 1$, in which concentrations are uncorrelated ($\mathrm{Cov}(\mathbf{y}, \mathbf{u}) = 0$) and the pure profiles overlap ($\mathbf{z}^T\mathbf{v} \neq 0$). We omit the explicit calculation of the PLS loadings, simply noting that they are some linear combinations of both **z** and **v** (cf. Figure 2). In relatively few lines, we can show that O-PLS exactly separates the analyte from the background.

We first note that the first normalized loading weight is proportional to the pure analyte profile, $\mathbf{w}_1 = \mathbf{z}/\|\mathbf{z}\|$. This follows from the fact that the covariance vector,

$$\mathbf{k} = \frac{1}{n}\mathbf{y}^T\left(\mathbf{y}\mathbf{z}^T + \mathbf{u}\mathbf{v}^T\right) = \mathrm{Var}(\mathbf{y})\mathbf{z}^T$$

analogous to **K** in Eqn 1.4, is collinear with **z** and proportional to $\mathbf{w}_1$ by construction. The initial score contains a **u**-component,

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 = \left(\mathbf{y}\mathbf{z}^T + \mathbf{u}\mathbf{v}^T\right)\mathbf{z}/\|\mathbf{z}\| = \alpha\mathbf{y} + \beta\mathbf{u}$$

where, for convenience, we define $\alpha = \|\mathbf{z}\|$ and $\beta = \frac{\mathbf{z}^T\mathbf{v}}{\|\mathbf{z}\|}$. O-PLS next computes the residual,

$$\begin{aligned}\mathbf{E}_{xy} &= \mathbf{X} - \mathbf{t}_1\mathbf{w}_1^T \\ &= \mathbf{y}\mathbf{z}^T + \mathbf{u}\mathbf{v}^T - (\alpha\mathbf{y} + \beta\mathbf{u})\frac{\mathbf{z}^T}{\alpha} \\ &= \mathbf{u}\left(\mathbf{v}^T - \frac{\beta}{\alpha}\mathbf{z}^T\right)\end{aligned}$$

and orthogonal loadings weight,

$$\begin{aligned}\tilde{\mathbf{w}}_{\mathrm{orth}}^T = \mathbf{t}_1^T\mathbf{E}_{xy} &= (\alpha\mathbf{y} + \beta\mathbf{u})^T\mathbf{u}\left(\mathbf{v}^T - \frac{\beta}{\alpha}\mathbf{z}^T\right) \\ &= n\beta\,\mathrm{Var}(\mathbf{u})\left(\mathbf{v}^T - \frac{\beta}{\alpha}\mathbf{z}^T\right)\end{aligned}$$

where the tilde simply indicates that the vector is not yet normalized. Note that $\mathbf{w}_{\mathrm{orth}}$ is proportional to the **z**-orthogonal projection of **v** because

$$\mathbf{v} - \frac{\beta}{\alpha}\mathbf{z} = \mathbf{v} - \frac{\mathbf{v}^T\mathbf{z}}{\|\mathbf{z}\|^2}\mathbf{z} = \mathbf{v} - (\mathbf{v}^T\check{\mathbf{z}})\check{\mathbf{z}}$$

where $\check{\mathbf{z}}$ is the unit vector in the direction of **z**.

Now let $\mathbf{w}_{\mathrm{orth}} = \tilde{\mathbf{w}}_{\mathrm{orth}}/\|\tilde{\mathbf{w}}_{\mathrm{orth}}\|$, where $\|\tilde{\mathbf{w}}_{\mathrm{orth}}\| = n\beta\,\mathrm{Var}(\mathbf{u})(\|\mathbf{v}\|^2 - \mathbf{v}^T\check{\mathbf{z}}^2)$. The fact that $\mathbf{w}_{\mathrm{orth}}$ is orthogonal to **z** implies that the orthogonal score reduces to exactly **u** via

$$\mathbf{t}_{\mathrm{orth}} = \mathbf{X}\mathbf{w}_{\mathrm{orth}} = \mathbf{u}\frac{n\beta\,\mathrm{Var}(\mathbf{u})\left(\|\mathbf{v}\|^2 - \mathbf{v}^T\check{\mathbf{z}}^2\right)}{\tilde{\mathbf{w}}_{\mathrm{orth}}} = \mathbf{u}$$

This leads to the first and only orthogonal loading being exactly equal to the background profile by

$$\mathbf{p}_{\mathrm{orth}}^T = \mathbf{t}_{\mathrm{orth}}^T X/\|\mathbf{t}_{\mathrm{orth}}\|^2 = \mathbf{u}^T\mathbf{u}\mathbf{v}^T/\|\mathbf{u}\|^2 = \mathbf{v}^T$$

It follows immediately that the background component is removed because $\mathbf{t}_{\mathrm{orth}}\mathbf{p}_{\mathrm{orth}}^T = \mathbf{u}\mathbf{v}^T$ and that the filtered data are exactly equal to the **y**-component, $\mathbf{X}_1 = \mathbf{y}\mathbf{z}^T$. O-PLS recomputes the first score,

$$\mathbf{t}_1 = \mathbf{X}_1\mathbf{w}_1 = \frac{\mathbf{y}\mathbf{z}^T\mathbf{z}}{\|\mathbf{z}\|} = \|\mathbf{z}\|\mathbf{y}$$

with corresponding (scalar) **y**-loading,

$$c_1 = \frac{\mathbf{t}_1^T\mathbf{y}}{\|\mathbf{t}_1^2\|} = \frac{\|\mathbf{z}\|\mathbf{y}^T\mathbf{y}}{\|\mathbf{z}\|^2\mathbf{y}^T\mathbf{y}} = \frac{1}{\|\mathbf{z}\|}$$

and regression vector,

$$\beta_{\text{O-PLS}} = c_1\mathbf{w}_1 = \frac{\mathbf{z}}{\|\mathbf{z}\|^2} = \frac{\check{\mathbf{z}}}{\|\mathbf{z}\|}$$

which is also proportional to the loading **p**. Although not derived here, we note that PLS also arrives at the same regression coefficient.

In summary, O-PLS filters exactly the **u**-component of the data, resulting in an easily interpreted, single-component model. Although PLS arrives at the same regression coefficient, each of the three scores and loadings depend significantly on the

degree of overlap of the pure profiles. This example contains only two constituents for simplicity, but analogous results hold for the example in Figure 2 and should generalize easily to more complex systems.

2.4.1. *Relation to baseline drift*

A common type of **Y**-orthogonal measurement error is baseline drift, whereby the mean of the sample spectra appears to vary for each sample [5]. We discuss in Section 3 how it is, in fact, possible for O-PLS not to filter this type of variation for certain types of data matrices, although the conditions under which this can happen are not likely to be found in calibration data. We ignore this possibility for the time being and show explicitly, for calibration data at least, why O-PLS identifies the drift as **Y**-orthogonal.

A useful way to think of baseline drift is as noise uncorrelated with the calibration concentrations but whose loading is a nonzero constant. As such, the noise "pure profile" may be thought of as overlapping with the constituent pure profiles and identified as **Y**-orthogonal. To see this in a simple case, replace the background component in the previous example with a "one" of the form $\sigma\xi^T$, where σ is a normally distributed n-vector with mean zero and unit variance (uncorrelated with **y**) and $\xi = \xi 1$ is a constant "baseline" vector, equal to a scalar, ξ, times the one-vector, 1. Thus the drift plays the role of the background in the example and is shown to be filtered accordingly.

2.5. Partial least squares prediction optimality for noise-free data and relation to orthogonal projection to latent structures

In this section, we provide justification for the observation that, in practice, O-PLS prediction is very similar to that of PLS. This means that, under the conditions in Section 2.4, O-PLS may have fewer components although being just as predictive as PLS. We begin by citing a relevant result about PLS prediction optimality by Nadler and Coifman [14]. We then show that PLS regression in the sense of Eqn 1.1 is identical whether one uses **X** or a filtered **X**, provided that the filtering is performed in a certain way. In particular, PLS prediction does not depend on the component of **z** that is orthogonal to the net analyte signal (NAS) vector. This suggests a mechanism by which O-PLS filters the data without impairing prediction.

Following the analysis by Nadler and Coifman [14], consider a single-response, noise-free system obeying Beer's law (Eqn 2.3 with $j = 1$ and $k \geq 1$). We assume that the $(k+1)\times(k+1)$ covariance matrix of $\{\mathbf{y}, \mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is of full rank and that $\mathbf{z} \notin \mathrm{Span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$. The analysis also depends on the notion of a NAS vector of the response, **y**, which is commonly defined as the component of the analyte profile that is orthogonal to the background profiles [16,17],

$$\mathbf{z}_{\mathrm{NAS}}(\mathbf{y}) = \mathbf{z} - \sum_{i=1}^{d_k} (\mathbf{z}^T \check{\mathbf{v}}_i)\check{\mathbf{v}}_i$$

Here, the set $\{\check{\mathbf{v}}_i\}_{i=1}^{d_k}$ is an orthonormal basis for $\mathrm{Span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$. Note that, by construction, the NAS is orthogonal to all of the background profiles,

$$\mathbf{z}_{\mathrm{NAS}}^T \mathbf{V} = 0 \qquad (2.11)$$

Nadler and Coifman show that, for an infinite training set with no noise, the following two results hold:

(1) the root mean square error of prediction (RMSEP) of a PLS model with at most $k+1$ components is zero, and
(2) the corresponding regression vector β_{PLS} is equal to a constant b times the NAS vector, that is, $\beta_{\mathrm{PLS}} = b\mathbf{z}_{\mathrm{NAS}}$.

Combining these two results, we find

$$\hat{\mathbf{y}} = \mathbf{X}\beta_{\mathrm{PLS}} = (\mathbf{y}\mathbf{z}^T + \mathbf{U}\mathbf{V}^T)\beta_{\mathrm{PLS}} = \mathbf{y}\mathbf{z}^T\beta_{\mathrm{PLS}} \qquad (2.12)$$

by Eqn 2.11 and result 2. This shows that exact prediction does not depend on the k background constituents. Under the stated conditions, any filtering procedure that does not alter the **y**-component of the data does not influence prediction optimality of PLS regression.

In fact, letting $\mathbf{z}_{\mathrm{NAS}}^{\perp} = \mathbf{z} - \mathbf{z}_{\mathrm{NAS}}$, we continue the derivation in Eqn 2.12 to find

$$\mathbf{X}\beta_{\mathrm{PLS}} = \mathbf{y}(\mathbf{z}_{\mathrm{NAS}} + \mathbf{z}_{\mathrm{NAS}}^{\perp})^T\beta_{\mathrm{PLS}} = (\mathbf{y}\mathbf{z}_{\mathrm{NAS}}^T)\beta_{\mathrm{PLS}}$$

From this, we conclude that any algorithm filtering **X** of its NAS-orthogonal components will not affect prediction optimality under the given conditions.

These observations do not make reference to an explicit O-PLS model for the most general case of correlated concentrations and overlapping pure profiles, but they do suggest an explanation for the similarity of PLS and O-PLS predictions. As long as filtering preserves the **y**-component completely, or the NAS-collinear part of **y**, regression on the filtered data yields identical prediction. And this "prediction agnostic" filtering is exactly the type that O-PLS is expected to perform. The question of how exactly O-PLS predictions are affected by the presence of noise for finite training data remains open. Nadler and Coifman [15] derive several interesting PLS results for noisy, finite training data that presumably could be extended to O-PLS.

2.6. General calibration

For the most general calibration data, it may not be appropriate to assume that analyte and background concentrations are uncorrelated. Here, we remove that assumption ($\mathrm{Cov}(\mathbf{Y}, \mathbf{U}) \neq 0$) and place no constraints on the pure profile relationships ($\mathbf{Z}^T\mathbf{V} \neq 0$). We demonstrate that filtering the variation can adversely affect interpretability of the resulting model. Furthermore, we show that it is likely in this case that no **Y**-orthogonal variation exists at all, although, intuitively, one might expect it to exist.

A simple example is sufficient to illustrate how filtering of **Y**-orthogonal variation in the most general case can adversely affect interpretation of the O-PLS model. Consider a three-component system with one analyte characterized by concentration **y** and pure profile **z**, and two background constituents characterized by concentrations $\mathbf{u}_1$, $\mathbf{u}_2$ and spectral profiles $\mathbf{v}_1$, $\mathbf{v}_2$ (Figure 3). We assume that **y** is correlated with $\mathbf{u}_1$ but not with $\mathbf{u}_2$ and that $\mathbf{u}_1$ and $\mathbf{u}_2$ are only partially correlated. In addition, all three pure profiles overlap in the middle region of the spectrum, corresponding to the case $a_T = 0.316$ in Figure 2. Intuitively, one might expect it to be beneficial to remove one orthogonal component because $\mathbf{u}_2$ is uncorrelated with **y**. Indeed, the removal of one orthogonal component results in a two-component model, where each score is more collinear with **y** than

any of the three PLS scores. One can see in the figure that the orthogonal loading positively weights the background peaks. However, removing this component degrades the interpretation of the model because the peaks of $\mathbf{v}_1$ are positively weighted in the regression vector. In contrast, the PLS regression vector shows positive weights for peaks in the analyte profile, $\mathbf{z}$, and negative weights for the background peaks.

Intuitively, the reason for poor interpretability of the regression coefficient is that O-PLS removes the $\mathbf{y}$-orthogonal component of both background peaks simultaneously, whereas only one of the background components is truly uncorrelated with the analyte. We find by inspection that the orthogonal loading $\mathbf{p}_{\text{orth}}$ is approximately equal to the first loading of the truly $\mathbf{y}$-orthogonal component of $\mathbf{X}$, denoted $\mathbf{X}^{\perp}$,

$$\mathbf{X}^{\perp} \approx \mathbf{u}_1^{\perp}\mathbf{v}_1^T + \mathbf{u}_2^{\perp}\mathbf{v}_2^T$$

where $\mathbf{u}_1^{\perp}$ and $\mathbf{u}_2^{\perp}$ are the $\mathbf{y}$-orthogonal components of $\mathbf{u}_1$ and $\mathbf{u}_2$. Hence, the filtered data are approximately equal to

$$\mathbf{X}_1 \approx \mathbf{X} - \mathbf{X}^{\perp} = \mathbf{y}\left(\mathbf{z}^T + \gamma\mathbf{v}_1^T\right)$$

where $\gamma = \dfrac{\text{Cov}(\mathbf{u}_1, \mathbf{y})}{\text{Var}(\mathbf{y})}$. It is as if the system now contains just the analyte, but with a modified profile, $\mathbf{z}^T + \gamma\mathbf{v}_1^T$. This is apparent in the figure, where one can see that the regression coefficient has significant contributions from $\mathbf{v}_1$ and essentially no interference from $\mathbf{v}_2$. Although it is true that the peaks related to $\mathbf{y}$ and $\mathbf{u}_1$ covary in the spectral data on account of correlated concentrations, they are not generally useful in predicting $\mathbf{y}$.

Under the mentioned conditions, it is also possible that no identifiable $\mathbf{Y}$-orthogonal variation exists. Consider the general covariance matrix for both PLS and O-PLS,

$$\begin{aligned}\mathbf{K} &= \frac{1}{n}\left(\mathbf{Y}^T\mathbf{Y}\mathbf{Z}^T + \mathbf{Y}^T\mathbf{U}\mathbf{V}^T\right) \\ &= \text{Cov}(\mathbf{Y})\mathbf{Z}^T + \text{Cov}(\mathbf{Y}, \mathbf{U})\mathbf{V}^T\end{aligned} \tag{2.13}$$

Recall that, by Eqn 1.4, the loading weights, $\mathbf{W}$, are the PCA loadings of $\mathbf{K}$. From Eqn 2.13, it is clear that these loading weights have contributions from the background constituent profiles, the significance of which depends directly on the analyte–background covariance. If $\mathbf{K}$ is of full rank, say with rank r_k, and the subspace spanned by the pure profiles (both calibration and background) is of dimension less than or equal to r_k, then $\mathbf{W}$ will contain an orthonormal basis for the row space of $\mathbf{X}$. It follows that $\mathbf{E}_{xy} = 0$ and O-PLS will find no orthogonal components. We explicitly outline properties of the data matrices that can lead to this situation in Section 3.1.

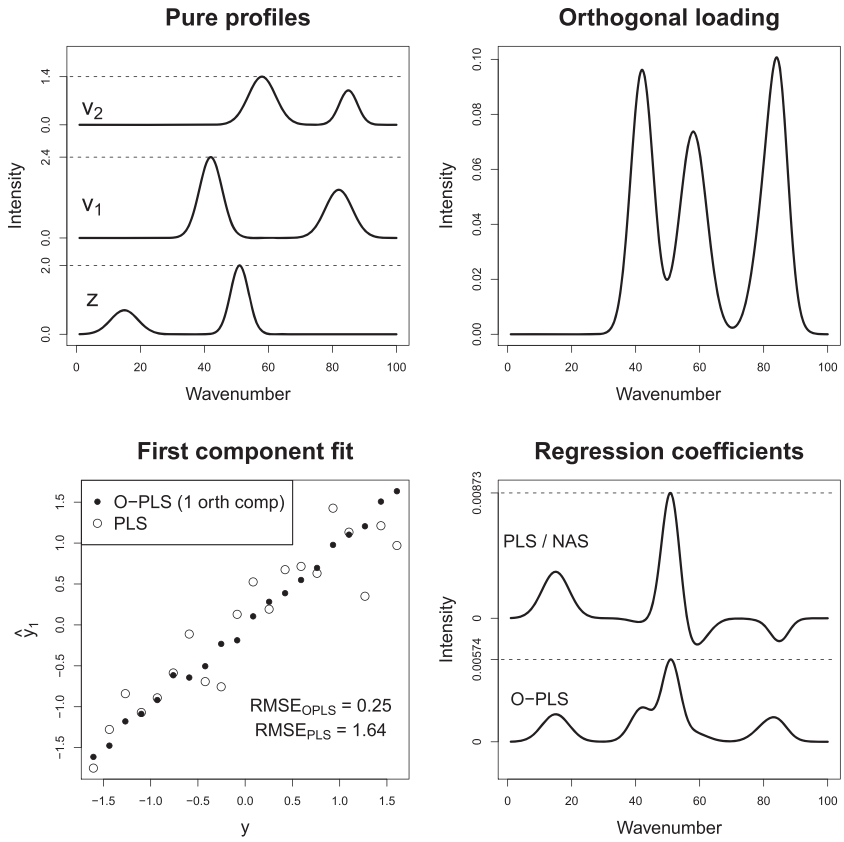

Figure 3. Comparison of PLS and O-PLS solution properties for system with overlapping pure profiles and concentrations (not shown) satisfying Cov ($\mathbf{y}$, $\mathbf{u}_1$) = 0.45, Cov($\mathbf{y}$, $\mathbf{u}_2$) = 0, and Cov($\mathbf{u}_1$, $\mathbf{u}_2$) = 0.04. Note that $\mathbf{u}_2$ is totally uncorrelated with $\mathbf{y}$ and has low correlation with $\mathbf{u}_2$, but its pure spectrum overlaps with both. Top left: Pure analyte profile ($\mathbf{z}$) and background pure profiles ($\mathbf{v}_1$ and $\mathbf{v}_2$). Top right: First O-PLS orthogonal loading. Bottom left: Comparison of fit for first O-PLS and PLS components. Bottom right: Comparison of O-PLS and PLS regression coefficients. Despite the orthogonality of $\mathbf{u}_2$ and $\mathbf{y}$, removal of one orthogonal component leads to a misleading regression coefficient that positively weights the peaks of $\mathbf{v}_1$, which, in general, are not useful for predicting $\mathbf{y}$.

3. ORTHOGONAL PROJECTION TO LATENT STRUCTURES FOR SYSTEMS BIOLOGY

3.1. Y-orthogonal variation may not exist

So far we have considered the performance of O-PLS on calibration data modeled by Beer's law. The benefit of having such a model is that it allows explicit analysis of solution properties and prediction capability. Although O-PLS interpretability can suffer when analyte and background concentrations are highly correlated, the mentioned analysis shows that it is otherwise quite effective at reducing the required number of predictive components.

Perhaps not surprisingly, in light of the success of O-PLS in chemometrics applications, recent attempts have been made to apply O-PLS to large two-block data sets in the context of systems biology. For large matrices, however, the data may not contain the type of orthogonal variation that the algorithm is designed to filter. We outline some fairly broad conditions that lead to this situation. By broad, we mean conditions that may be commonly encountered when the number of variables in **X** and **Y** is large relative to the number of samples, as might be the case in studies involving microarray or metabolomic data with anywhere from hundreds to tens of thousands of variables. Barring successful filtering of structured **Y**-orthogonal variation, the model produced by O-PLS is identical to a PLS model of the original data. We note that, in contrast to the previous section, the analysis of this section does not depend on assumptions of large n or the absence of noise.

Before presenting technical details, we take a general view of how O-PLS may encounter problems when both data matrices have a large number of variables. Suppose both **X** and **Y** contain far more variables than observations ($m, p >> n$). The purpose of both PLS and O-PLS is to find linear combinations of the independent **X**-variables that simultaneously describe variation in **X** and are significant for the prediction of **Y**. Recall that O-PLS filters portions of the scores that are orthogonal to the columns of **Y**, whereas PLS scores may have **Y**-orthogonal components. However, when **Y** is *large enough* and *enough* of the response variables are linearly independent, the subspace orthogonal to the columns of **Y** is empty. In other words, because each column of **Y** is a vector in the vector space R^n spanned by an n-dimensional basis set, if enough of the columns are linearly independent, then all n of those basis vectors will be needed to construct the best possible predictive scores, leaving no basis vectors to describe a possibly **Y**-orthogonal subspace. This can lead to situations where, intuitively, orthogonal variation exists but the algorithm is unable to characterize it.

3.1.1. Conditions on X and Y

We now describe the technical properties of **X** and **Y** that can lead to a situation in which O-PLS is unable to identify **Y**-orthogonal variation. A less general form of the conditions is presented first, both because it is easier to understand and because it is likely to occur in real data. For simplicity, we assume that the mean-centered data are of full rank, although all of the arguments may be easily modified in the event that mean centering reduces the matrix ranks by one. Proofs are available in the Appendix.

Theorem

(a) Let $\mathbf{X} \in R^{n \times p}$ *and* $\mathbf{Y} \in R^{n \times m}$ *be two data matrices containing n samples of p independent and m response variables, respectively. If* $m, p \geq n$, *and that* rank(**X**) = rank(**Y**) = n *with all n components of PCA*(**K**) *considered significant for O-PLS, then* $\mathbf{E}_{xy} = 0$ *and O-PLS fails to find* **Y***-orthogonal variation.*

A more general statement is the following.

Theorem

(b) If colspace(**X**)⊆colspace(**Y**) *and* $r_x = \text{rank}(\mathbf{X})$ *components of PCA*(**K**) *are considered significant, then* $\mathbf{E}_{xy} = 0$ *and O-PLS fails to find* **Y***-orthogonal variation.*

We note that matrices satisfying the conditions in (a) automatically satisfy those in (b), so (a) is technically a corollary of (b). To see this, suppose the conditions in (a) are satisfied. Then colspace(**X**) = colspace(**Y**) = R^n because both matrices are of full rank and $r_x = n$ components of PCA(**K**) are considered significant.

An important question for the application of O-PLS is how often can one expect to encounter the given conditions in real data. In calibration studies with a relatively small number of analytes, it is unlikely that one will encounter either set of conditions. In particular, with $m < n \leq p$, O-PLS can potentially find orthogonal variation. The conditions in (a) are impossible by default because $\text{rank}(\mathbf{Y}) < n$. And in order for the conditions in (b) to be true, the spectra would have to be extremely simple, with all relevant **X**-variables being described by linear combinations of the measured concentrations. Except for the simplest data, this is unlikely to happen.

The situation is quite different, however, when the number of response variables is greater than the number of samples, which is especially true if one is trying to correlate, for example, transcriptomic data with metabolomic data [9]. In this case, both matrices will almost certainly be of maximal rank n. The question then becomes whether or not the covariance matrix **K** will also be of rank n. Heuristically, this can happen when n or more significant factors change independently across the samples. Under these conditions, O-PLS may be unable to identify systematic variation.

In some sense, looking for orthogonal variation when the feature space is exceptionally large appears to be fundamentally inappropriate. If, instead, we start with the assumption that a relatively small number of factors are cross-correlated in an interesting way, we open the door for a host of other successful approaches. For example, one may try to enforce sparsity of the regression coefficients using a shrinkage-selection algorithm [18] or, as in canonical correlation analysis, attempt to identify linear combinations of the features in **X** and **Y** simultaneously that best describe the correlations between them [19].

3.2. A simple example

We illustrate with an example in which there is no **Y**-orthogonal variation in **X**, yet there are aspects of **X** that one's intuition says should be filtered. Imagine a study in which five strains of a microorganism each has a unique genetic signature, resulting in a compound that is apparent in its Mid-IR spectrum. The aim of the study is to try to uncover information about both the genetic pathways involved and the corresponding changes in cell chemistry. Now suppose the data show that exactly one distinct gene in the microarray is expressed and that the expression is highly correlated with one distinct peak in the Mid-IR spectrum (Figure 4). If we assume mean centering of the data, then **X**, **Y**, and the covariance matrix **K** are of maximal rank n. Because each gene is independent of (and hence uncorrelated with) the

others and has significant correlations with one peak in the spectrum, one cannot discard loadings of **K** without discarding essential information about the quantities of interest. As a result, a full *n*-component model is needed.

Now suppose the samples in this experiment have been contaminated, with the result that a spurious peak appears in the spectral measurements. Although we may intuitively expect this peak to be uncorrelated with gene expression, the given statements about the rank of **X**, **Y**, and **K** are still valid. Hence, the conditions in Theorem (a) are met and O-PLS will not filter the variation. In this case, PLS and O-PLS models are identical and contain significant contributions from the spurious peak.

One way to circumvent the inability of O-PLS to identify **Y**-orthogonal variation is to increase the number of observations. Geometrically, the five genes plus the systematic variation in our small example are six independent factors that the algorithm is attempting to describe by scores in a five-dimensional subspace. Increasing the sample size, then, not only results in more accurate sample estimates to covariance (collinearity) and orthogonality (uncorrelatedness) between the variables, but also increases the dimensionality of the vector subspace used to describe true variation in the data. Indeed, if two observations of each experimental state are taken instead of one, O-PLS coefficients are seen to correspond nicely with the correct spectral peaks, and the orthogonal loading (not shown) corresponds directly to the spurious peak.

Another way to avoid this issue is to force O-PLS to identify **Y**-orthogonal variation by building a separate model for each of the genes. In the example, for instance, this leads to the spurious peak being identified as orthogonal to each $\{\mathbf{y}_i\}_{i=1}^{j}$. Unfortunately, one is left with *m* separate models of the same variance in **X**, as well as up to *m* independent sets of latent variables. If the type of orthogonal variation is known *a priori*, one could try to use O-PLS in a somewhat surgical way via categorical response variables designed to target specific system variation [20]. This approach is not general, however, because it depends on a fairly specific knowledge of the confounding effects in the system.

4. CONCLUSION

For spectral data obeying Beer's law in the noise-free, large *n* setting, O-PLS often generates regression models that are as good at prediction as PLS, but with far fewer components. The extent to which O-PLS separates analyte and background constituents depends primarily on the covariance structure between the two groups. At its best, when the concentrations of interest are uncorrelated with the background, O-PLS loadings and regression coefficients are linear combinations of the pure profiles. Otherwise, although prediction scores remain collinear with the calibration concentrations, interpretability of loadings and regression coefficients is degraded because they contain contributions from the background profiles. The extent of the degradation depends in a complex way on concentration covariance and spectral interference of the pure profiles.

For noise-free and error-free data, PLS prediction is known to be optimal with respect to RMSEP. However, components of **X** uncorrelated with the calibration concentrations or, more generally, those whose loadings are orthogonal to the NAS vector do not affect the optimality of PLS prediction. Because O-PLS explicitly

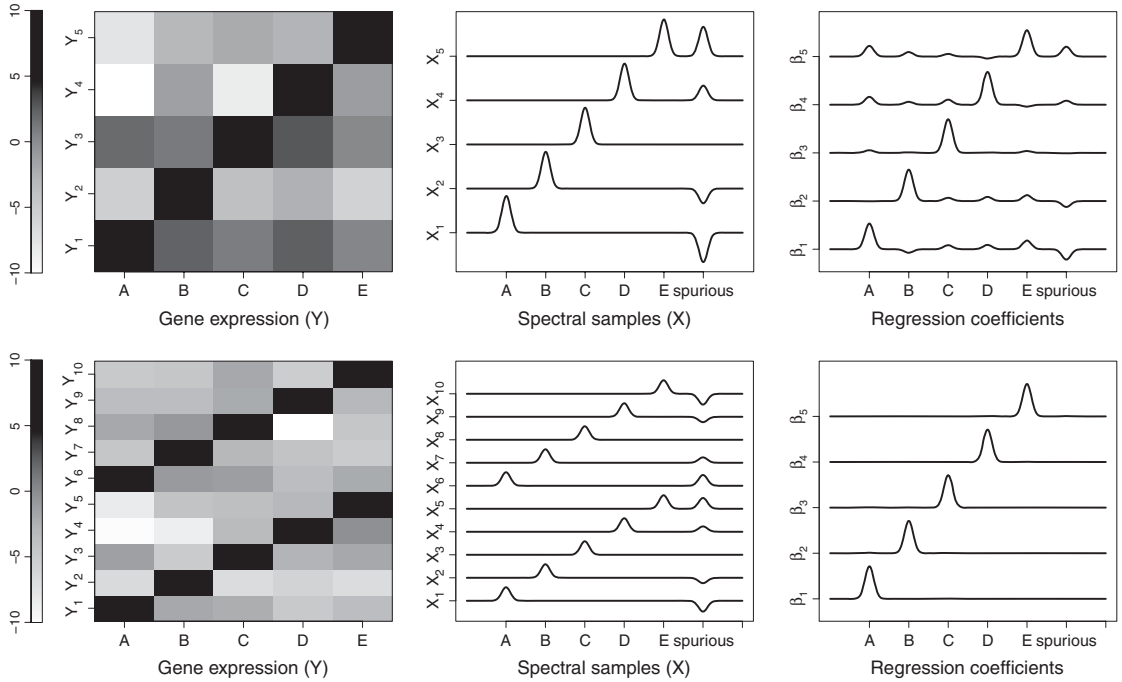

Figure 4. An illustration of how O-PLS ability to filter orthogonal variation depends on properties of the data matrices. Top panel: Data consisting of only $n = 5$ samples of gene expression **Y** (left) are assumed to have a spurious peak on the right side of the measured spectra **X** (center) that is truly uncorrelated with gene expression. The microarray heat maps show gene expression, increasing in value from black to white. Each gene has a distinct signature, as indicated by the corresponding letters on the *x*-axis. Solid lines show the true spectra, whereas dashed lines indicate systematic variation that is uncorrelated with **Y**. O-PLS does not identify orthogonal variation, and the regression coefficients β (right) are adversely affected by the systematic variation. Bottom panel: If each measurement is carried out twice, the orthogonality of the spurious peak becomes more apparent (left). O-PLS can then identify and filter the spectral data (center) such that the orthogonal variation is removed (right).

filters scores that are **Y**-orthogonal, this suggests a reason for the similarity of prediction between the two algorithms.

Despite the success of the algorithm in the context of calibration studies, caution is needed when applying O-PLS to data sets with large feature spaces (for which both $p, m \geq n$), a situation that may be commonly encountered in the context of systems biology. In this scenario, it is likely that O-PLS will be unable to identify **Y**-orthogonal variation as defined by the algorithm. The issue may be remedied by increasing the number of samples. Although doing so in chemometrics may be relatively quick and inexpensive, the same cannot be said for certain types of experiments in systems biology (e.g., microarray) in which the number of observations may be unavoidably small. If prediction is the aim, identification of **Y**-orthogonal variation may not be crucial, although the lack of theoretical models to describe the data makes evaluation of prediction quality completely dependent on data-driven methods.

If O-PLS or PLS are instead used for the exploration of features in large systems biology data sets, the algorithms are likely to lead to factorizations of the data that are difficult to interpret. Preprocessing by clustering or variable selection may improve the situation by giving the algorithms smaller, more manageable subsets on which to work. However, the search for single vectors explaining maximal variance tends to conflate intuitively independent features. Depending on the application, it may be beneficial to instead incorporate specific knowledge directly into whichever algorithm is being used, be it non-negativity of the spectral matrices [21,22], statistical independence of the underlying factors [23,24], or sparsity of the data [18,25].

Acknowledgements

We acknowledge the contributions of Kofi Adragni, Jinsuk Lee, and Terry Haut for many helpful discussions, as well as Ryan Elmore for a critical review of the manuscript. This research was supported the DOE Office of Biological and Environmental Research, grant award no. DE-AC36-08GO28308, and by the BioEnergy Research Center. The BioEnergy Research Center is a US Department of Energy BioEnergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Additional support was provided by a grant from the Department of Energy, Office of Science, SciDAC, and GTL programs and by the ASCR and BioEnergy Research Offices within the Office of Science, grant award no. DE-AC36-99GO10337.

REFERENCES

1. Geladi P, Kowalski BR. *Anal. Chim. Acta* 1986; **185**: 1–17.
2. Martens H, Naes T. *Multivariate Calibration*. Wiley: Chichester 1989.
3. Wold S, Sjostrom M, Eriksson, L. *Chemometr. Intell. Lab.* 2001; **58**: 109–130.
4. Laurens LML, Wolfrum EJ. *Bioenerg. Res.* 2011; **4**(1): 22–35.
5. Trygg J, Wold S. *J. Chemom.* 2002; **16**: 119–128.
6. Rantalainen M, Bylesjo M, Cloarec O, Nicholson JK, Holmes E, Trygg J. *J. Chemom.* 2007; **21**: 376–385.
7. Trygg J, Wold S. *J. Chemom.* 2003; **17**: 53–64.
8. Bylesjo M, Eriksson D, Kusano M, Moritz T, Trygg J. *Plant J.* 2007; **52**: 1181–1191.
9. Bylesjo M, Nilsson R, Srivastava V, Gronlund A, Johansson AI, Jansson S, Karlsson J, Moritz T, Wingsle G, Trygg J. *J. Proteome Res.* 2009; **8**: 199–210.
10. Kemsley EK, Tapp HS. *J. Chemom.* 2009; **23**: 263–264.
11. Ergon R. *J. Chemom.* 2005; **19**: 1–4.
12. Ergon R. *J. Chemom.* 2007; **21**: 537–546.
13. Verron T, Sabatier R, Joffre R. *J. Chemom.* 2004; **18**: 62–68.
14. Nadler B, Coifman RR. *J. Chemom.* 2005; **19**: 45–54.
15. Nadler B, Coifman RR. *J. Chemometr.* 2005; **19**: 107–118.
16. Bro R, Andersen CM. *J. Chemom.* 2003; **17**: 646–652.
17. Lorber A, Faber K, Kowalski BR. *Anal. Chem.* 1997; **69**: 1620–1626.
18. Tibshirani R. *J. R. Stat. Soc., B* 1996; **58**: 267–288.
19. Hardoon DR, Szedmak S, Shawe-Taylo J. *Neural Comput.* 2004; **16**: 2639–2664.
20. Bylesjo M, Eriksson D, Sjodin A, Jansson S, Moritz T, Trygg J. *BMC Bioinform.* 2007; **8**: 207.
21. Paatero P. *Chemometr. Intell. Lab.* 1997; **37**: 23–35.
22. Lee DD, Seung HS. *Nature* 1999; **401**: 788–791.
23. Hyvarinen A, Oja E. *Neural Netw.* 2000; **13**: 411–430.
24. Liebermeister W. *Bioinformatics* 2002; **18**: 51–60.
25. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd edn). Springer-Verlag: New York 2009.

Appendix

The following are the proofs of Theorems (a) and (b).

Proof

(a) Because both **X** and **Y** are assumed to be of rank n and $m, p \geq n$, **K** is of rank n and, by assumption that all PCA components of **K** are significant, we can write $\mathbf{K} = \mathbf{C}\mathbf{W}^T$ from Eqn 1.4 with $A = n$. A standard result of linear algebra is that the dimension of the row space of **X** is equal to the matrix rank n, so we just need to show that each column of **W** belongs to the row space of **X**. Then the columns of **W** must form an orthonormal basis for the row space of **X** and the result follows.

We show that the columns of **W** belong to the row space of **X** by contradiction. Assume that column i of **W** is orthogonal to the row space of **X**. Multiplying both sides of Eqn 1.4 by $\mathbf{X}^T$ from the right, we see on the right hand side that $\mathbf{X}\mathbf{X}^T$ is an $n \times n$ matrix of rank n, implying that $\mathbf{K}\mathbf{X}^T = \mathbf{Y}^T\mathbf{X}\mathbf{X}^T$ is also of rank n. On the left hand side, the $n \times n$ matrix $\mathbf{W}^T\mathbf{X}^T$ is of rank at most n-1 because row i is composed of all zeros (by the orthogonality of row i of **W** to all rows of **X**). This implies $\mathbf{K}\mathbf{X}^T = \mathbf{C}\mathbf{W}^T\mathbf{X}^T$ is of rank at most $nn1$, a contradiction. Therefore, all of the columns of **W** belong to the row space of **X** and, because they are orthonormal and rank(**X**)= n, form an orthonormal basis for the space. Hence, $\mathbf{X}\mathbf{W}\mathbf{W}^T = \mathbf{X}$ and $\mathbf{E}_{xy} = 0$.

(b) The assumption that the column space of **X** belongs to the column space of **Y** implies that $r_y = \text{rank}(\mathbf{Y}) \geq r_x$, which in turn implies that $\text{rank}(\mathbf{K}) = r_x$. Because all r_x of these components are assumed to be significant, the loading matrix **W** contains r_x orthonormal basis vectors in its columns. On account of the fact that $dim(rowspace(\mathbf{X})) = dim(colspace(\mathbf{X})) = r_x$, the columns of **W** form an orthonormal basis for the row space of **X**. The proof of this is identical to the one given (i.e., assume at least one column of **W** is orthogonal to the row space of **X**, and show that this leads to a contradiction about the rank of $\mathbf{K}\mathbf{X}^T$). It follows again that $\mathbf{X}\mathbf{W}\mathbf{W}^T = \mathbf{X}$ and $\mathbf{E}_{xy} = 0$. □