Subspace Correction Methods in Multivariate Calibration

THÈSE N^{O} 4919 (2011)

PRÉSENTÉE LE 4 FÉVRIER 2011

À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE D'AUTOMATIQUE

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Pamandeep Singh GUJRAL

acceptée sur proposition du jury:

Prof. R. Longchamp, président du jury
Prof. D. Bonvin, Dr M. Amrhein, directeurs de thèse
Prof. S. Morgenthaler, rapporteur
Dr M. Rhiel, rapporteur
Prof. J. Trygg, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

To my grandfather

Acknowledgments

I owe my deepest gratitude to Professor Dominique Bonvin and Dr Michael Amrhein for their indispensable support and mentorship during the course of this work. While Michael's insightful project proposal opened up the PhD position that has lead to this dissertation, Professor Bonvin provided his perspective and assistance in developing a roadmap to achieve the stated objectives. Michael's affable nature, unbeatable enthusiasm, and absolute devotion to research made it a pleasure to work under his guidance. His invaluable contribution to the preparation and editing of this text is gratefully acknowledged.

I warmly thank Dr Michal Dabros for his collaboration on many research projects, that facilitated my smooth transition from electronics to chemometrics. This research also benefited tremendously from helpful suggestions of Dr Barry M. Wise during his sabbatical at the LA. Thanks go to Professor Bala Srinivasan and Professor Rolf Ergon for insightful discussions; to Professor Rasmus Bro, Professor Martin Hassler, and Professor Wulfram Gerstner for their inspiring courses; to the president of the defense jury, Professor Roland Longchamp, and the external jury members Professor Johan Trygg, Professor Stephan Morgenthaler, and Dr Martin Rhiel for reviewing this dissertation. This work was partially supported by a financial grant from the Swiss National Science Foundation, through project number 200021-109582.

The realization of this dissertation would not have been a real fulfillment without the good times with friends and colleagues. Thanks are due to Nirav, whose desk has been my one stop shop for career counseling, LaTeX support,

and an inexhaustible supply of Indian snacks at all odd (non)-working hours. I will fondly remember the golden moments of my Swiss journey with Alejandro, Micheal, and Carla; and the gyaan distribution over lunch get-togethers with Willson and Saurabh. Many thanks especially to Bjoern and Sandeep for their engagement in various sport activities, and to Gorka, Evgeny, Alain, Sean, Mark, Marc, and Seb, who continue to inspire me with their untiring zeal. I extend my gratitude to the friendly administrative staff - Ruth, Francine, Sara, Homeira, Francis, Philippe Cuanillon - for their sincere support on numerous occasions. I appreciate the efforts of Philippe, Greg, Sandy, Andriana, David, Laleh, and Lina for organizing all the fun activities at the LA, and congratulate all the chefs for the excellent management of everything from pepper inventory to orgy celebrations. The truly wonderful LA atmosphere will be greatly missed.

Finally, I offer my most heartfelt thanks to my grandmother, parents, brother, and sister for their endless love and constant encouragement.

Abstract

Productivity, quality, safety, and environmental concerns have driven major advancements in the development of process analyzers. Analyzers generate measurement data that are useful for characterizing product and process attributes (key variables), thereby benefiting the drive towards automatic control and optimization. However, these objectives may be severely compromised when key variables are determined at low sampling rates through off-line analysis. It is sometimes possible to relate more easily available secondary measurements (predictors) to key variables (predictands) using data-driven soft sensors or calibration models. These models can then be used to deliver information about key variables at a higher sampling rate and/or at lower financial burden.

This work studies multivariate calibration for spectroscopic measurements (such as near-infrared, mid-infrared, ultra-violet, Raman spectra, or nuclear magnetic resonance) that are linked to concentrations of one or more analytes using an inverse regression model based on principal component regression (PCR) or partial least-squares regression (PLSR). Spectroscopic measurements are typically corrupted with both random zero-mean measurement errors (*noise*) and systematic variations (*drift*) caused by instrumental, operational and process changes. The prediction error can be decomposed into the error due to noise in the calibration data and bias resulting from truncation in PCR/PLSR, and the error due to drift and noise in the prediction data. To correct for these errors, this work proposes three subspace correction methods that use new information in addition to calibration data. Firstly, latent subspace correction using *unlabeled data* (secondary measurements for which the key variables are unknown)

helps reduce the error due to noise in the calibration data and truncation. Secondly, drift subspace correction is achieved following a two-step procedure. In the first step, the drift subspace is estimated using *slave data* with drift and *master data* with no drift. In the second step, the original calibration data are corrected for the estimated drift subspace using shrinkage or orthogonal projection. The third subspace correction method involves data reconciliation, which is the procedure of adjusting predicted key variables to obtain estimates that are consistent with balance equations. The various methodologies are illustrated using both simulated and experimental data.

Keywords: spectroscopic measurements; latent variables; unlabeled data; systematic disturbances; drift; shrinkage; orthogonal projection; data reconciliation.

Résumé

La poductivité, la qualité, la sécurité, et les préoccupations environnementales conduisent à des progrès majeurs dans le développement d'analyseurs de processus. Ceux-ci génèrent des données de mesure qui sont utiles pour caractériser les attributs des produits et processus (variables clés), bénéficiant ainsi au développement des applications en réglage automatique ou en optimisation. Cependant, ces objectifs peuvent être gravement compromis lorsque les variables clés sont déterminées à un faible taux d'échantillonnage, lors de l'analyse hors-ligne par exemple. Toutefois, il est parfois possible de corréler des variables secondaires (prédicteurs) aux variables clés (prédicats) en utilisant des capteurs logiciels utilisant les données et des modèles d'étalonnage. Ces modèles peuvent ensuite être utilisés pour fournir des informations concernant les variables clés à taux d'échantillonnage plus élevé avec ou sans coût supplémentaire.

Le présent travail étudie l'étalonnage multivariable pour les mesures spectroscopiques (que sont par exemple, l'infrarouge proche, l'infrarouge moyen, l'ultra-violet, les spectres de Raman, ou la résonance magnétique nucléaire). Dans notre cas, ces mesures sont liées à des concentrations d'un ou plusieurs analystes à travers un modèle de régression inverse basé soit sur la méthodes des composantes principales (PCR), soit sur la méthodes des moindres carrés partiels (PLSR). Les mesures spectroscopiques sont généralement entachés d'erreurs de mesure aléatoires de moyenne nulle (*bruit*), et les variations systématiques (*dérive*) sont causées par la variabilité des instruments de mesure, par les conditions opérationnelles et par la variations du processus. L'erreur de prédiction peut être décomposée en l'erreur due au bruit dans les données

d'étalonnage et de distorsion résultant de la troncature en PCR/PLSR, et celle due à la dérive et le bruit dans les données de prévision. Afin de corriger ces erreurs, nous proposons trois méthodes de sous-espace de correction qui utilisent l'information supplémentaire par rapport aux données d'étalonnage.

Tout d'abord, une correction fondée sur un sous-espace latent utilisent des données non étiquetées (ce sont des mesures secondaires pour lesquelles les variables clés demeurent inconnues). Ceci améliore l'erreur provoquée par le bruit dans les données d'étalonnage. Deuxièmement, une méthode de compensation de dérive utilisant un sous-espace adapté est proposée qui se décompose en deux étapes. Durant la première étape, le sous-espace de dérive est estimé en utilisant des *données esclaves* comportant une dérive et des *données de référence* qui ne comportent pas de dérive. A la second étape, les données d'étalonnage originales sont corrigées à partir du sous-espace de dérive à l'aide de mises à l'échelle et de projections orthogonales. La troisième méthode fondée sur les sous-espaces de correction se fonde sur la réconciliation de données. Elle ajuste les variables clés prédites afin d'obtenir des estimées qui sont consistantes avec les équations de bilan. Les diverses méthodes sont illustrées à travers à la fois des données de simulation et des données expérimentales.

Mots-clés : mesures spectroscopiques ; variables latentes ; données non étiquetées ; perturbations systématiques ; dérive ; retrait ; projection orthogonale, réconciliation de données.

Contents

List of Abbreviations

API	active pharmaceutical ingredient
ASTM	American Society for Testing and Materials
CC	component correction
CLS	classical least squares
DCPS	difference correction of prediction spectra
DOP	dynamic orthogonal projection
DR	data reconciliation
ECM	explicit correction method
EPO	external parameter orthogonalization
EROS	error removal by orthogonal subtraction
FDA	Food and Drug Association
FTIR	Fourier-transform infrared
GLS	generalized least squares
ICM	implicit correction method
iid	independent and identically distributed
IIR	independent interference reduction
IT-PCR	PCR proposed by Isaksson and Thomas
LV	latent variable
MIR	mid-infrared
MLR	multiple linear regression
MSC	multiplicative scatter correction

NAS	net-analyte-signal
NASR	net-analyte-signal regression
NIPALS	nonlinear iterative partial least squares
NIR	near-infrared
NMR	nuclear magnetic resonance
NSO-PLSR	non-sequential optimized PLSR
OF	optimal filtering
OLS	ordinary least squares
OP	orthogonal projection
OPLS	orthogonal partial least squares
OSC	orthogonal signal correction
PACLS	prediction-augmented classical least squares
PAT	process analytical technology
PCA	principal component analysis
PCA-PLSR	PCA-based PLSR
PCR	principal component regression
PGNA	prompt gamma neutron activation
pNAS	pseudo net-analyte-signal
pNASR	pseudo net-analyte-signal regression
PLSR	partial least-squares regression
PSD	positive semi-definite
QbD	quality by design
RRMSEP	relative root-mean-squared error of prediction
SNR	signal-to-noise ratio
SNV	standard normal variate
SO-PLSR	sequential optimized PLSR
SVD	singular value decomposition
SVM	support vector machine
SVR	support vector regression
TOP	calibration transfer by orthogonal projection
UV	ultraviolet
XRD	X-ray diffraction
XRF	X-ray fluorescence

1

Introduction

1.1 Overview

Productivity, quality, safety, and environmental concerns have driven major advancements in the development of process analyzers for food, (bio-)chemical, pharmaceutical, petrochemical, paper/pulp, cement, steel, semiconductor, and related process industries. The analyzers generate measurement data that are useful for characterizing product and process attributes (so-called key variables), thereby benefiting the drive towards automatic control and optimization. In accordance with the philosophy of "Quality by Design" (QbD)- *Quality cannot be tested into products; quality should be built in by design* - the process analytical technology (PAT) initiative by the US Food Drug Administration (FDA) encourages the pharmaceutical industries to design process measurement systems to allow controlling manufacturing with the goal of ensuring final product quality [1]. However, these objectives may be severely compromised when key variables are determined at low sampling rates through off-line laboratory analysis. It is sometimes possible to relate more easily available secondary measurements[1] to key variables[2] using data-driven soft sensors or calibration models, which can be used to deliver information about key variables at a higher sampling rate and/or at lower financial burden [2]. For example, in pharmaceutical manufacturing of tablets, a set of scalar measurements – such as temperature, pH, pressure, flowrates, calorimetry, turbidity, CO_2 concentration, and vector

[1] *predictors*, or X-measurements
[2] *predictands*, y-measurements, properties of interest, or response variables

measurements – such as spectrum, chromatogram, particle size or shape distribution, can be related to the key variables such as tablet homogeneity, dissolution rate, humidity, active pharmaceutical ingredient (API) concentration, impurities, polymorphism and crystallographic characteristics [3].

This work studies latent variable (LV) calibration models, with spectroscopy as a prototypical example. Spectrometers are used routinely to measure the intensities of light absorbed, reflected, or emitted by a sample to determine its physical or chemical make up. Due to their many advantages – small sampling times, easy in-situ installation, low maintenance requirements, inherent sterility, non-invasiveness and non-destructiveness – spectrometers are widely employed in research and industry. Spectroscopic measurements from near-infrared (NIR), mid-infrared (MIR), ultra-violet (UV), Raman, nuclear magnetic resonance (NMR), X-ray fluorescence (XRF), X-ray diffraction (XRD), or prompt gamma neutron activation (PGNA), can be linked linearly to molar concentrations or weight percentages of an analyte of interest. Often, multivariate spectroscopic calibration involves highly correlated or collinear predictors, with the number of predictors greatly outnumbering the sample size. Together, the observations (as rows) and predictor variables (as columns) can be represented as a 'data matrix' $\mathbf{X}$ (see Fig. 1.1).

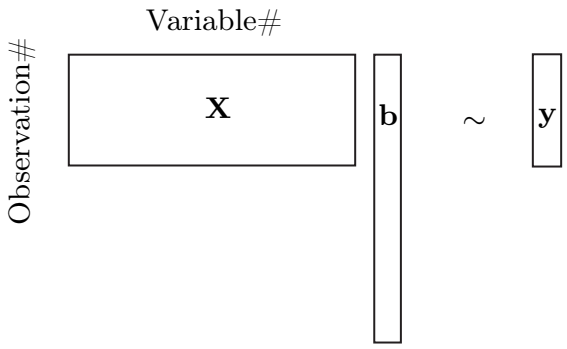


Fig. 1.1. Schematic of data matrix $\mathbf{X}$, key variables $\mathbf{y}$, and the regression vector $\mathbf{b}$, such that $\mathbf{y} \sim \mathbf{Xb}$.

Estimation of the regression vector $\mathbf{b}$ becomes difficult in high-dimensional spaces due to the increasing sparseness of data. With the number of variables greater than the number of observations, the problem is ill-posed. Standard methods such as ordinary least squares (OLS) are indeterminate and hence can-

not be applied, while multiple linear regression (MLR) leads to overfit models with little predictive power [4]. This necessitates some form of regularization. A wealth of methods are available in the statistics and the machine learning literature such as ridge regression [5], locally-weighted regression [6], least absolute shrinkage and selection operator [7], projection pursuit regression [8], multivariate adaptive regression splines [9], artificial neural networks [10], support vector regression [11], Gaussian process regression [12], science-based calibration [13]. Subspace regression models, such as classical least squares (CLS), principal component regression (PCR), and partial least-squares regression (PLSR) [14], solve the problem of high dimensionality by projecting the data onto a lower-dimensional subspace (*latent subspace*), resulting in a set of fewer predictor variables (*latent variables*), that can be regressed onto the key variables $\mathbf{y}$. CLS, PCR, and PLSR differ in how the latent subspace is defined. While CLS is a forward calibration procedure that requires knowledge of all analytes in the system (which is rarely the case), PCR/PLSR are inverse calibration models that require knowledge of only the analyte of interest [15]. The subject matter of this dissertation is limited to PCR and PLSR, that are the *de facto* standard calibration methods in spectroscopy [15, 16].

Though the focus of this dissertation is on multivariate spectroscopic calibration, the same principles apply to any vector-based measurement that is linked linearly to some underlying latent quantity or response variable. For the measurement vector $\mathbf{x}$, the response variable y, and the LVs $\mathbf{t}$, the difference between spectroscopic calibration and the LV framework is shown in Fig. 1.2. In spectroscopic calibration (see Fig. 1.2i), $\mathbf{x}$ has a causal relationship[3] with y, and $\mathbf{t}$ is an hypothetical construct or a mathematical quantity that is correlated with both $\mathbf{x}$ and y. On the other hand, the LV framework includes the case shown in Fig. 1.2ii, where both $\mathbf{x}$ and y are caused by $\mathbf{t}$, hence allowing one to build a model to predict y from $\mathbf{x}$. From the point of view of predictive learning, the two cases in Fig. 1.2 are equivalent. Furthermore, the scope of this dissertation also includes classification models that proceed via regression, i.e. where PCR or PLSR are used to find a low-rank representation of the data, followed by any linear or nonlinear discriminant function.

[3] A change in y causes a change in $\mathbf{x}$

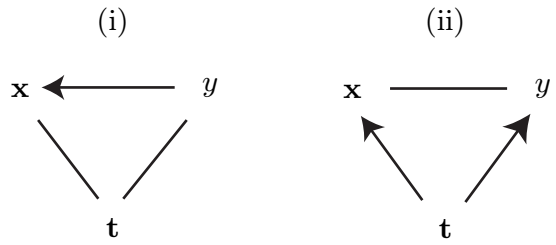


Fig. 1.2. Schematic of (i) spectroscopic calibration, and (ii) LV framework. Directed edges indicate causal relationship, while undirected edges indicate correlation.

1.2 Outline of the thesis

The present dissertation is composed of six chapters. Fig. 1.3 shows the outline of the thesis with the aid of a simplified flowchart. An overview of the material studied is given below, while the state-of-the-art and the proposed methods are described in the relevant chapters.

Chapter 2 presents the fundamentals of LV calibration based on PCR and PLSR, and the specific attributes of calibration with spectroscopic measurements. Most often, calibration data consist of *labeled data* only, i.e. the X-measurements for which the corresponding y-measurements are available. Based on this estimated calibration model, the y-values are predicted from new X-measurements.

In regression modeling, it is generally assumed that the labeled data are representative of the prediction data. However, in the prediction step, the X-measurements may be corrupted by systematic variations (*drift*) caused by instrumental, operational and process changes. If the new X-measurements have *unseen drift* [4], the original calibration model typically results in large prediction errors. **Chapter 3** discusses a two-step framework of methods for *drift subspace correction*. In the first step, the drift subspace is estimated based on calibration data and small amounts of so-called *master/slave data* (e.g. X-measurements at standard room temperature can be considered as master data, and X-measurements of the same samples at ± 5 °C can be considered as slave data). In the second step, the part of the calibration data lying in the estimated

[4] Drift is decomposed into drift seen in the calibration data and drift unseen in the calibration data. Unseen drift is more problematic.

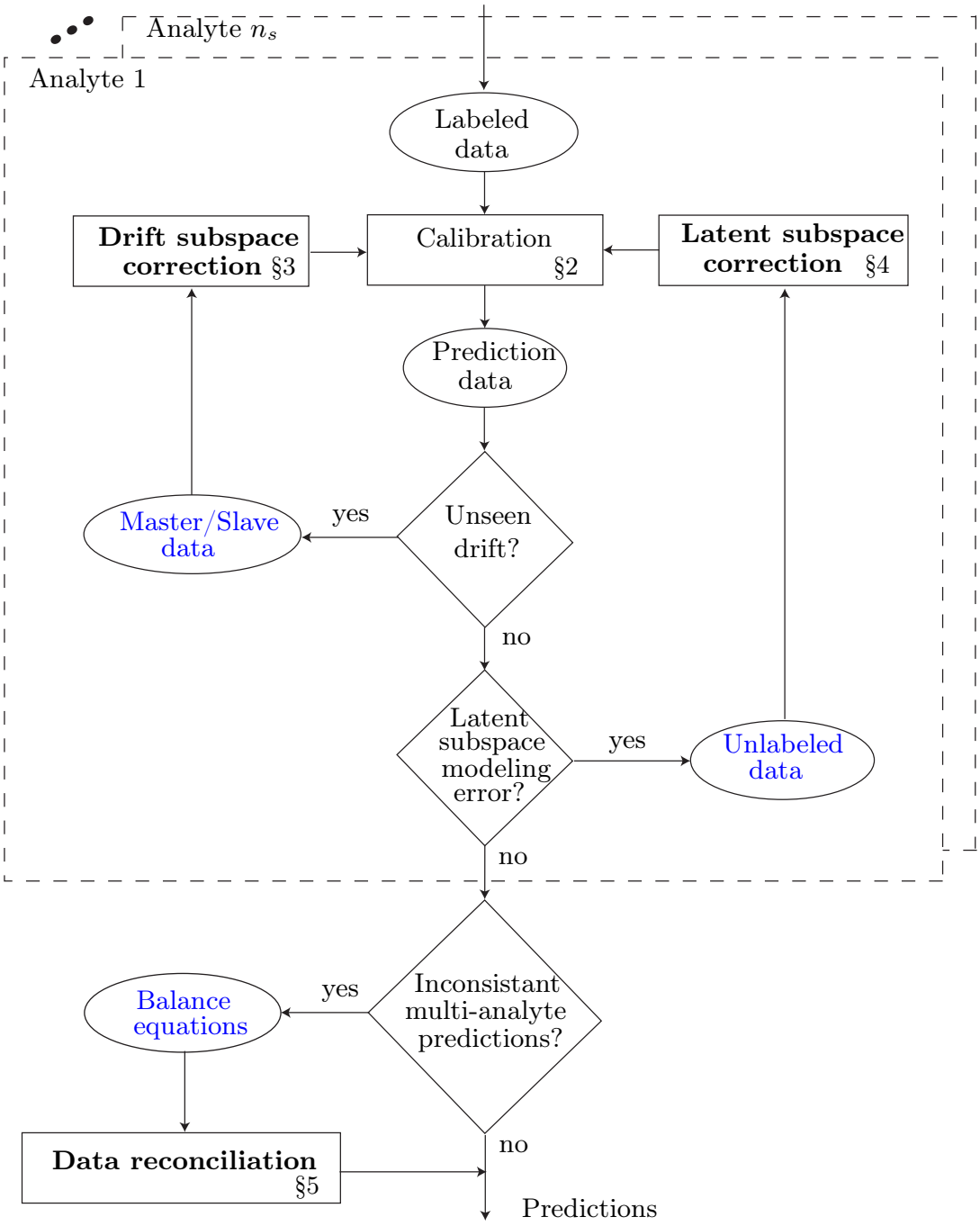


Fig. 1.3. Flowchart showing the different components of this dissertation work and the corresponding chapter numbers. A source of (new) information is shown in an ellipse, working procedure in a rectangle, and a decision step in a rhombus.

drift subspace is scaled down. Re-calibration with the scaled data results in a calibration model that is less sensitive to drift directions.

In many practical applications, the labeled data are small owing to the high costs of reference measurements. However, a large number of X-measurements may be available at-line/off-line (so-called *unlabeled data*), which may be used together with the labeled data for calibration. The use of unlabeled data in addition to labeled data helps stabilize the latent subspaces in the calibration step, typically leading to lower subspace modeling errors. *Latent subspace correction* based on the use of unlabeled data is discussed in **Chapter 4**. It is shown that, prediction data, which by definition qualifies as unlabeled data, can be used for latent subspace correction only if it has been corrected for unseen drift.

Chapter 5 studies a third form of subspace correction, that is applicable when calibration models are built for several analytes. Linear dependencies may exist between the analytes due to e.g. chemical reactions. These balance equations can be used to adjust and improve the predictions in a *data reconciliation* (DR) step. **Chapter 6** draws some general conclusions from the work.

2

Preliminaries

In this chapter, some background information is given that will be used in the thesis.

2.1 Notations

The following notations will be used throughout the dissertation. Matrices are represented by bold capital letters (e.g. $\mathbf{X}$), column vectors by bold lower-case letters (e.g. $\mathbf{y}$), and scalars by lower-case letters in italics (e.g. α). $\mathbf{1}_n$, $\mathbf{0}_n$, $\mathbf{I}_n$, and $\mathbf{0}_{[n\times q]}$ denote an n-dimensional vector of ones, zeros, the $[n \times n]$ identity matrix and the $[n \times q]$ matrix of zeros, respectively. The transpose operator is represented by $(.)^{\mathrm{T}}$, the Moore-Penrose inverse by $(.)^{+}$, the l_2-norm of a vector by $||.||$, the expectation operator by $\mathcal{E}$, the span of row vectors and the column vectors of a matrix by $\mathcal{R}(.)$ and $\mathcal{C}(.)$, respectively, and the normal distribution with mean μ and standard deviation σ by $\mathcal{N}(\mu, \sigma)$. Measured quantities are indicated with $\tilde{(.)}$, and estimated quantities with $\hat{(.)}$. Subscripts are introduced where necessary to distinguish between variables of the same type (e.g. prediction or calibration data).

2.2 LV calibration

2.2.1 Overview of LV framework

Let $\tilde{\mathbf{x}}$ denote the n_x-dimensional measurement vector (with measurement noise $\mathbf{v}_x$) from a first-order instrument, and $\tilde{y}$ the measured response variable (with measurement noise v_y). In the LV framework with r factors, the model for $\tilde{\mathbf{x}}$ and $\tilde{y}$ is written as:

$$\begin{aligned} \tilde{\mathbf{x}} = &\quad \mathbf{x} + \mathbf{v}_x = \mathbf{L}\,\mathbf{t} + \mathbf{v}_x \\ \tilde{y} = &\quad y + v_y = \mathbf{q}^{\mathrm{T}}\mathbf{t} + v_y \,, \end{aligned} \tag{2.1}$$

where $\mathbf{t}$ is an r-dimensional vector of scores (or LVs), $\mathbf{L}$ an $[n_x \times r]$ orthogonal X-loading matrix that defines the underlying low-dimensional space in which the noise-free X-measurements lie, and $\mathbf{q}$ the r-dimensional y-loading vector. For n_c calibration measurements (*labeled data*), Eq. (2.1) can be written in matrix form as:

$$\begin{aligned} \tilde{\mathbf{X}}_c = &\quad \mathbf{X}_c + \mathbf{V}_{x,c} = \quad \mathbf{T}_c\mathbf{L}^{\mathrm{T}} + \mathbf{V}_{x,c} \\ \tilde{\mathbf{y}}_c = &\quad \mathbf{y}_c + \mathbf{v}_{y,c} = \quad \mathbf{T}_c\mathbf{q} + \mathbf{v}_{y,c} \,, \end{aligned} \tag{2.2}$$

where $\tilde{\mathbf{X}}_c$ is the $[n_c \times n_x]$ X-measurement matrix, $\tilde{\mathbf{y}}_c$ the n_c-dimensional y-measurement vector, $\mathbf{T}_c$ the $[n_c \times r]$ scores matrix, $\mathbf{V}_{x,c}$ the $[n_c \times n_x]$ X-noise matrix, and $\mathbf{v}_{y,c}$ an n_c-dimensional vector of y-noise. The linear regression model reads:

$$\tilde{\mathbf{y}}_c = \tilde{\mathbf{X}}_c\mathbf{b} + \mathbf{f}_c \,, \tag{2.3}$$

where $\mathbf{b}$ is the n_x-dimensional regression vector, and $\mathbf{f}_c$ the n_c-dimensional vector of residuals. The regression vector $\mathbf{b}$ can be estimated from the data pair $\{\tilde{\mathbf{X}}_c, \tilde{\mathbf{y}}_c\}$ using standard regression methods such as PCR and PLSR that are described in the following sections.

2.2.2 PCR

Before discussing PCR, it is useful to introduce two useful tools from matrix algebra, the singular value decomposition (SVD) and principal component analysis (PCA). The SVD of the $[n_c \times n_x]$ matrix $\tilde{\mathbf{X}}_c$ is written as:

$$\tilde{\mathbf{X}}_c = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^{\mathrm{T}} \,, \tag{2.4}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $[n_c \times n_x]$ and $[n_x \times n_x]$ orthonormal matrices, and $\mathbf{\Sigma}$ is an $[n_x \times n_x]$ diagonal matrix with the singular values of $\tilde{\mathbf{X}}_c$ in decreasing order on its diagonal.

The scores and loadings based on the PCA factorization of $\tilde{\mathbf{X}}_c$ and retaining r_{PCR} factors is written as:

$$\tilde{\mathbf{X}}_c = \mathbf{T}_{\text{PCA}}\,\mathbf{P}_{\text{PCA}}^{\text{T}} + \mathbf{E}_{\text{PCA}}\,, \tag{2.5}$$

where $\mathbf{T}_{\text{PCA}}$ is an orthogonal scores matrix of dimension $[n_c \times r_{\text{PCR}}]$, $\mathbf{P}_{\text{PCA}}$ is an orthonormal loading matrix of dimension $[n_x \times r_{\text{PCR}}]$, and $\mathbf{E}_{\text{PCA}}$ contains the residues that are orthogonal to both the scores and the loading matrices, i.e. $\mathbf{E}_{\text{PCA}}\mathbf{P}_{\text{PCA}} = \mathbf{0}_{n_c \times r_{\text{PCR}}}$ and $\mathbf{T}_{\text{PCA}}^{\text{T}}\mathbf{E}_{\text{PCA}} = \mathbf{0}_{r_{\text{PCR}} \times n_x}$. Let $\mathbf{U}_r$ and $\mathbf{V}_r$ contain only the first r_{PCR} columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and $\mathbf{\Sigma}_r$ be the upper left $[r_{\text{PCR}} \times r_{\text{PCR}}]$ portion of $\mathbf{\Sigma}$. Then, $\mathbf{T}_{\text{PCA}} = \mathbf{U}_r\,\mathbf{\Sigma}_r$ and $\mathbf{P}_{\text{PCA}} = \mathbf{V}_r$.

The first step of PCR involves the SVD or PCA factorization, retaining r_{PCR} factors. In the second step, the regression vector $\hat{\mathbf{b}}$ is computed from the least-squares regression between $\{\mathbf{T}_{\text{PCA}}, \tilde{\mathbf{y}}_c\}$:

$$\begin{aligned} \mathbf{q}_{\text{PCR}} &= \mathbf{T}_{\text{PCA}}^{+}\tilde{\mathbf{y}}_c \\ \hat{\mathbf{b}} &= \mathbf{P}_{\text{PCR}}\mathbf{q}_{\text{PCR}}\,, \end{aligned} \tag{2.6}$$

where $\mathbf{q}_{\text{PCR}}$ is the estimated r_{PCR}-dimensional y-loading vector. In contrast to OLS or MLR, that perform least squares with a large number of variables $\left(= \text{rank}\left(\tilde{\mathbf{X}}_c\right)\right)$, PCR performs least squares with r_{PCR} (for collinear data r_{PCR} may be $\ll n_c, n_x$) scores, thereby regularizing $\hat{\mathbf{b}}$. The regularization in PCR can be seen from the following equivalent description of the optimization done in PCR:

$$\begin{aligned} \hat{\mathbf{b}} &= \mathit{argmin}||\tilde{\mathbf{X}}_c\mathbf{b} - \tilde{\mathbf{y}}_c|| \quad \text{such that} \quad \mathbf{b} \in \mathcal{C}(\mathbf{P}_{\text{PCR}}) \\ &= \mathit{argmin}||\tilde{\mathbf{X}}_c\,\mathbf{P}_{\text{PCR}}\mathbf{P}_{\text{PCR}}^{\text{T}}\,\mathbf{b} - \tilde{\mathbf{y}}_c||\,, \end{aligned} \tag{2.7}$$

i.e. $\hat{\mathbf{b}}$ is constrained to lie in the space defining the directions of maximum variation in $\tilde{\mathbf{X}}_c$, thereby making $\hat{\mathbf{b}}$ less sensitive to noise.

2.2.3 PLSR

Two different algorithms for PLSR were developed in parallel [16]. Wold *et al.* developed a solution with orthogonal score vectors and non-orthogonal loading vectors and, for that purpose, they used the nonlinear iterative partial least squares (NIPALS) algorithm [17]. Martens, on the other hand, developed an algorithm resulting in non-orthogonal score vectors and orthogonal loading vectors. As shown by Ergon [18], one model can be converted into the other. The two PLSR models are presented next, followed by a discussion on the similarities and differences in the two models.

Wold's PLSR using $\{\tilde{\mathbf{X}}_c, \tilde{\mathbf{y}}_c\}$ and retaining r_{PLSR} factors is established following the NIPALS algorithm:

1. Let $\mathbf{X}_0 = \tilde{\mathbf{X}}_c$. For $a = 1, \ldots, r_{\text{PLSR}}$ perform Steps 2–6.
2. Compute $\mathbf{w}_a = \mathbf{X}_{a-1}^{\text{T}} \tilde{\mathbf{y}}_c / ||\mathbf{X}_{a-1}^{\text{T}} \tilde{\mathbf{y}}_c||$
3. Compute $\mathbf{t}_a = \mathbf{X}_{a-1} \mathbf{w}_a$
4. Compute $q_a = \tilde{\mathbf{y}}_c^{\text{T}} \mathbf{t}_a (\mathbf{t}_a^{\text{T}} \mathbf{t}_a)^{-1}$
5. Compute $\mathbf{p}_a = \mathbf{X}_{a-1}^{\text{T}} \mathbf{t}_a (\mathbf{t}_a^{\text{T}} \mathbf{t}_a)^{-1}$
6. Deflate $\mathbf{X}_a = \mathbf{X}_{a-1} - \mathbf{t}_a \mathbf{p}_a^{\text{T}}$

Note that the deflation of $\mathbf{X}$ matrix in Step 6 can equivalently be defined with respect to the $\mathbf{y}$ vector. The resulting factorization is written as:

$$\begin{aligned} \tilde{\mathbf{X}}_c &= \mathbf{T}_{\text{W}} \mathbf{P}^{\text{T}} + \mathbf{E}_{\text{W}} \\ \tilde{\mathbf{y}}_c &= \mathbf{T}_{\text{W}} \mathbf{q}_{\text{W}} + \mathbf{f}\,, \end{aligned} \tag{2.8}$$

where the $[n_c \times r_{\text{PLSR}}]$ matrix $\mathbf{T}_{\text{W}} = [\mathbf{t}_1\, \mathbf{t}_2\, \ldots\, \mathbf{t}_{r_{\text{PLSR}}}]$ is orthogonal, $[n_x \times r_{\text{PLSR}}]$ matrix $\mathbf{P} = [\mathbf{p}_1\, \mathbf{p}_2\, \ldots\, \mathbf{p}_{r_{\text{PLSR}}}]$, r_{PLSR}-dimensional vector $\mathbf{q}_{\text{W}} = [q_1\, q_2\, \ldots\, q_{r_{\text{PLSR}}}]^{\text{T}}$ and the $[n_c \times n_x]$ matrix $\mathbf{E}_{\text{W}} = \mathbf{X}_{r_{\text{PLSR}}}$. The $[n_x \times r_{\text{PLSR}}]$ matrix $\mathbf{W} = [\mathbf{w}_1\, \mathbf{w}_2\, \ldots\, \mathbf{w}_{r_{\text{PLSR}}}]$ resulting from the algorithm is orthonormal. Furthermore, $\mathbf{T}_{\text{W}} = \tilde{\mathbf{X}}_c \mathbf{W} (\mathbf{P}^{\text{T}} \mathbf{W})^{-1}$ and $\mathbf{q}_{\text{W}} = (\mathbf{T}_{\text{W}}^{\text{T}} \mathbf{T}_{\text{W}})^{-1} \mathbf{T}_{\text{W}}^{\text{T}} \tilde{\mathbf{y}}_c$. The matrices $\mathbf{W}$ and $\mathbf{P}$ do not have a closed-form solution, their columns are obtained iteratively. The regression vector is computed as $\hat{\mathbf{b}} = \mathbf{W} (\mathbf{P}^{\text{T}} \mathbf{W})^{-1} \mathbf{q}_{\text{W}}$.

Martens proposed a different PLSR algorithm. However, Martens' PLSR model can be established once $\mathbf{W}$ is estimated using the NIPALS algorithm.

Retaining r_{PLSR} factors, the factorization used in Martens' PLSR [19, 20] is written as:

$$\begin{aligned} \tilde{\mathbf{X}}_c &= \mathbf{T}_{\text{M}}\mathbf{W}^{\text{T}} + \mathbf{E}_{\text{M}} \\ \tilde{\mathbf{y}}_c &= \mathbf{T}_{\text{M}}\mathbf{q}_{\text{M}} + \mathbf{f}\,, \end{aligned} \tag{2.9}$$

where $\mathbf{T}_{\text{M}} = \tilde{\mathbf{X}}_c\,\mathbf{W} = \mathbf{T}_{\text{W}}(\mathbf{P}^{\text{T}}\mathbf{W})$, and $\mathbf{q}_{\text{M}} = (\mathbf{T}_{\text{M}}^{\text{T}}\mathbf{T}_{\text{M}})^{-1}\mathbf{T}_{\text{M}}^{\text{T}}\tilde{\mathbf{y}}_c$. The regression vector is computed as $\hat{\mathbf{b}} = \mathbf{W}(\mathbf{W}^{\text{T}}\tilde{\mathbf{X}}_c^{\text{T}}\tilde{\mathbf{X}}_c\mathbf{W})^{-1}\mathbf{W}^{\text{T}}\tilde{\mathbf{X}}_c^{\text{T}}\tilde{\mathbf{y}}_c$.

Wold's PLSR and Martens' PLSR differ in how matrix deflations are carried out [20]. Wold's NIPALS algorithm orthogonalizes the columns of $\mathbf{X}$ and/or $\mathbf{y}$ with respect to the estimated scores, while Martens' PLSR orthogonalizes (the column of) $\mathbf{y}$ with respect to the scores and the rows of $\mathbf{X}$ with respect to $\mathbf{W}$. However, the two algorithms are equivalent in the sense that the score vectors span the same space $\mathcal{C}(\mathbf{T}_{\text{W}}) = \mathcal{C}(\mathbf{T}_{\text{M}})$, they result in the same regression vector $\hat{\mathbf{b}} = \mathbf{W}(\mathbf{P}^{\text{T}}\mathbf{W})^{-1}\mathbf{q}_{\text{W}} = \mathbf{W}(\mathbf{W}^{\text{T}}\tilde{\mathbf{X}}_c^{\text{T}}\tilde{\mathbf{X}}_c\mathbf{W})^{-1}\mathbf{W}^{\text{T}}\tilde{\mathbf{X}}_c^{\text{T}}\tilde{\mathbf{y}}_c$, and they result in the same y-residue $\mathbf{f}$ [19]. These similarities are tabulated in Table 2.1.

Similarities between Wold's and Martens' PLSR
same $\mathbf{W}$ and $\mathbf{f}$
$\mathcal{C}(\mathbf{T}_{\text{W}}) = \mathcal{C}(\mathbf{T}_{\text{M}})$
$\hat{\mathbf{b}} = \mathbf{W}(\mathbf{P}^{\text{T}}\mathbf{W})^{-1}\mathbf{q}_{\text{W}} = \mathbf{W}(\mathbf{W}^{\text{T}}\tilde{\mathbf{X}}_c^{\text{T}}\tilde{\mathbf{X}}_c\mathbf{W})^{-1}\mathbf{W}^{\text{T}}\tilde{\mathbf{X}}_c^{\text{T}}\tilde{\mathbf{y}}_c$

Table 2.1. Similarities between Wold's and Martens' PLSR.

In contrast to PCR, where the X-residuals are orthogonal to the scores, loadings, and regression vector[1], in Wold's PLSR $\mathbf{T}_{\text{W}}^{\text{T}}\mathbf{E}_{\text{W}} = \mathbf{0}_{r_{\text{PLSR}}\times n_x}$ but $\mathbf{E}_{\text{W}}\mathbf{P} \neq \mathbf{0}_{n_c\times r_{\text{PLSR}}}$ and $\mathbf{E}_{\text{W}}\hat{\mathbf{b}} \neq \mathbf{0}_{n_c}$, while in Martens' PLSR $\mathbf{E}_{\text{M}}\mathbf{W} = \mathbf{0}_{n_c\times r_{\text{PLSR}}}$ and $\mathbf{E}_{\text{M}}\hat{\mathbf{b}} = \mathbf{0}_{n_c}$ but $\mathbf{T}_{\text{M}}^{\text{T}}\mathbf{E}_{\text{M}} \neq \mathbf{0}_{r_{\text{PLSR}}\times n_x}$. The difference between the two PLSR factorizations stems from the definition of the model space, i.e. the space in which the relevant part of $\tilde{\mathbf{X}}_c$ lies. While Wold's PLSR uses oblique projections to define the relevant part of $\tilde{\mathbf{X}}_c$ as $\tilde{\mathbf{X}}_c\,\mathbf{W}(\mathbf{P}^{\text{T}}\mathbf{W})^{-1}\mathbf{P}^{\text{T}}$, Martens' PLSR uses orthogonal projections to define the same as $\tilde{\mathbf{X}}_c\,\mathbf{W}\mathbf{W}^{\text{T}}$. Different definitions

[1] Algebraic orthogonality is the geometric interpretation of statistical independence

of the model space results in different X-residuals, i.e. $\mathbf{E}_{\mathrm{W}} \neq \mathbf{E}_{\mathrm{M}}$. Hence, the choice between Wold's and Martens' PLSR may impact only if X-residuals are used e.g. for monitoring, fault diagnosis, or noise characterization.

Remarking that $\mathbf{T}_{\mathrm{M}} = \mathbf{T}_{\mathrm{W}}(\mathbf{P}^{\mathrm{T}}\mathbf{W})$ and $\mathbf{T}_{\mathrm{M}}\mathbf{q}_{\mathrm{M}} = \mathbf{T}_{\mathrm{W}}\mathbf{q}_{\mathrm{W}}$, Ergon proposed the following factorization [18]:

$$\begin{aligned} \tilde{\mathbf{X}}_c &= \mathbf{T}_{\mathrm{W}}\mathbf{P}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}} + \mathbf{E}_{\mathrm{M}} \\ \tilde{\mathbf{y}}_c &= \mathbf{T}_{\mathrm{W}}\mathbf{q}_{\mathrm{W}} + \mathbf{f}\,. \end{aligned} \tag{2.10}$$

The above factorization is essentially another form of Martens' PLSR as it retains the same X-residuals $\mathbf{E}_{\mathrm{M}}$, i.e. Ergon re-interpreted Martens' PLSR with scores $\mathbf{T}_{\mathrm{W}}$ and loadings $\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{P}$. The concept of PLSR as an optimal filter is explained in Chapter 5 (see also Appendix B). It is shown that Martens' and Ergon's PLSR lead to scores that are optimal with respect to the loadings and error covariance. However, Wold's PLSR leads to scores that are not optimal with respect to the loadings and error covariance.

As in the case of PCR, PLSR performs least squares with r_{PLSR} scores, thereby regularizing $\hat{\mathbf{b}}$. The regularization in PLSR can be seen from the following equivalent description of the optimization done in PLSR:

$$\begin{aligned} \hat{\mathbf{b}} &= argmin||\tilde{\mathbf{X}}_c\mathbf{b} - \tilde{\mathbf{y}}_c|| \quad \text{such that} \quad \mathbf{b} \in \mathcal{C}(\mathbf{W}) \\ &= argmin||\tilde{\mathbf{X}}_c\,\mathbf{W}\mathbf{W}^{\mathrm{T}}\,\mathbf{b} - \tilde{\mathbf{y}}_c||\,, \end{aligned} \tag{2.11}$$

i.e. $\hat{\mathbf{b}}$ lies in the lower-dimensional space of $\mathbf{W}$, thereby making $\hat{\mathbf{b}}$ less sensitive to noise. However, note that the constraint on $\hat{\mathbf{b}}$ does not follow from a global statistical optimization criterion, but is the outcome of an iterative algorithm.

2.2.4 Choice of meta-parameters r_{PCR} or r_{PLSR}

Comparing the LV model in (2.2), and the PCA and PLSR factorizations in Eqs. (2.5), (2.8), (2.9), and (2.10), one may be led to falsely assume that r_{PCR} and r_{PLSR} should be equal to the underlying dimensionality r (rank of $\mathbf{X}_c$ or the pseudo-rank of $\tilde{\mathbf{X}}_c$). However, since r_{PCR} and r_{PLSR} need to be optimized for prediction, and not for fitting $\tilde{\mathbf{X}}_c$, good choices of meta-parameters may not be the same as r [21]. The most commonly applied tool to aid the selection of meta-

parameters is cross-validation. In cross-validation, the relationship between prediction errors and meta-parameter values is estimated using calibration data. The following steps are performed:

1. Leave out part of the labeled data
2. Build the model without these data
3. Use the model to predict y-values on these data
4. Calculate the corresponding prediction error
5. Cycle through this procedure such that each data value has been left out at least once, and compute the mean of all squared prediction errors
6. Based on expert opinion, choose the number of factors that lead to "low" prediction errors, with "some" penalty on the number of factors chosen.

Note that different schemes exist to choose the part of the labeled data to be left out in Step 1. A related method is bootstrapping where, instead of building calibration models with different subsets of the labeled data, calibration models are built with subsamples of the labeled data [22]. Also note that Step 6 can only be guided by rules of thumb or heuristic criteria based on expert opinion as there exists no optimal strategy that is suitable for all kinds of data.

Cross-validation is often used to choose amongst different models (e.g. PCR or PLSR), different data pre-treatments (e.g. variable selection algorithms, mean-centering, smoothening, auto-scaling, normalization or standardization, scatter correction, alignment), and the model parameters (e.g. the number of factors retained in PCR/PLSR) and meta-parameters (e.g. width parameter and polynomial order in Savitsky-Golay smoothening) [15]. However, if the goal is to select the right combination of models, pre-treatments, and parameters from amongst a very large number of possible combinations, then cross-validation may lead to overfitting. Some extensions such as cross-model validation schemes, also known in the literature as two-fold or double cross-validation, have been proposed to reduce overfitting (see [23] and references within).

2.3 Calibration using spectroscopic data

2.3.1 Special case of LV calibration

Let $\tilde{\mathbf{x}}$ denote the spectroscopic measurement (absorbance) vector of an n_x-channel spectrometer, and $\mathbf{z}$ the n_s-dimensional concentration vector, where n_s is the number of absorbing species. For a spectroscopic measurement depending *linearly* on $\mathbf{z}$, e.g. when Beer-Lambert law is valid, one can write:

$$\tilde{\mathbf{x}} = \mathbf{S}\,\mathbf{z} + \mathbf{v}_x, \tag{2.12}$$

where $\mathbf{S}$ is the $[n_x \times n_s]$ pure-component spectra matrix consisting of the columns $\mathbf{s}_1, \ldots, \mathbf{s}_{n_s}$. Typically, a PCR/PLSR calibration model is built independently for each *species of interest* using labeled data alone. With $\tilde{y}$ as the measured concentration of the i^{th} species, and $\mathbf{L}\,\mathbf{t} = \mathbf{S}\,\mathbf{z}$, calibration with spectroscopic data can be cast into the LV framework. To avoid confusion, it is important to remark here that, throughout the thesis, $\mathbf{z}$ is used to refer to the concentration vector of n_s species, while $\mathbf{y}$ is used to refer to the concentration vector of one species over several samples.

For n_c calibration measurements, the X and y-measurements can be written in matrix form as:

$$\begin{aligned} \tilde{\mathbf{X}}_c &= \mathbf{Z}_c\mathbf{S}^{\mathrm{T}} + \mathbf{V}_{x,c} \\ \tilde{\mathbf{y}}_c &= \mathbf{y}_c + \mathbf{v}_{y,c} \\ \mathbf{y}_c &= i^{th} \text{ column of } \mathbf{Z}_c\,. \end{aligned} \tag{2.13}$$

where $\mathbf{Z}_c$ is of dimension $[n_c \times n_s]$, $\mathbf{V}_{x,c}$ and $\mathbf{v}_{y,c}$ are as defined in Eq. (2.2).

2.3.2 Net-analyte-signal

For noise-free X and y-measurements, PCR and PLSR lead to the same regression vector $\hat{\mathbf{b}} = (\mathbf{X}_c)^{+}\mathbf{y}_c$. Furthermore, if $\mathrm{rank}\,(\mathbf{S}) = \mathrm{rank}\,(\mathbf{Z}_c) = n_s$, $(\mathbf{X}_c)^{+}\mathbf{y}_c$ is the net-analyte-signal (NAS) vector [24]. The NAS vector for i^{th} species represents the portion of the pure-component spectrum of the i^{th} species that resides in the space orthogonal to the pure-component spectra of all other species. Let $\mathbf{Z}_{c,-i}$ and $\mathbf{S}_{-i}$ contain all but the i^{th} columns of $\mathbf{Z}_c$ and $\mathbf{S}$, respectively. Then $\hat{\mathbf{b}} \propto \mathbf{s}_{\mathrm{NAS,i}}$, where

$$\mathbf{s}_{\mathrm{NAS,i}} = (\mathbf{I}_{n_x} - (\mathbf{S}_{-i}^{\mathrm{T}})^{+}\mathbf{S}_{-i}^{\mathrm{T}})\,\mathbf{s}_i\,. \tag{2.14}$$

Since rank $(\mathbf{Z}_c) = n_s$ (or rank $(\mathbf{Z}_{c,-i}) = n_s - 1$), the right-hand-side of Eq. (2.14) projects $\mathbf{s}_i$ orthogonal to $\mathcal{R}(\mathbf{Z}_{c,-i}\mathbf{S}_{-i}^{\mathrm{T}})$. Intuitively, the NAS vector indicates a direction in which the X-measurements are affected only by changes in the concentration of the species of interest.

In order to relax the assumption of rank $(\mathbf{Z}_c) = n_s$, we define a pseudo-NAS (pNAS) vector for the i^{th} species of interest. Let rank $(\mathbf{S}) = n_s$, rank $(\mathbf{Z}_c) \leq n_s$, and rank $(\mathbf{Z}_{c,-i}) =$ rank $(\mathbf{Z}_c) - 1$. The pNAS vector $\mathbf{s}_{\mathrm{pNAS,i}}$ is defined:

$$\mathbf{s}_{\mathrm{pNAS,i}} := \left(\mathbf{I}_{n_x} - (\mathbf{Z}_{c,-i}^{+}\mathbf{Z}_{c,-i}\mathbf{S}_{-i}^{\mathrm{T}})^{+}(\mathbf{Z}_{c,-i}^{+}\mathbf{Z}_{c,-i}\mathbf{S}_{-i}^{\mathrm{T}})\right)\,\mathbf{s}_i\,. \tag{2.15}$$

Similar to the case of NAS, the right-hand-side of Eq. (2.15) projects $\mathbf{s}_i$ orthogonal to $\mathcal{R}(\mathbf{Z}_{c,-i}\mathbf{S}_{-i}^{\mathrm{T}})$.

The regression model with $\hat{\mathbf{b}} \propto \mathbf{s}_{\mathrm{NAS,i}}$ will be referred to as NAS regression (NASR), and with $\hat{\mathbf{b}} \propto \mathbf{s}_{\mathrm{pNAS,i}}$ as pseudo-NASR (pNASR). Note that, $\mathbf{S}$ and hence the NAS/pNAS vector are usually not known. Also, while the NAS/pNAS vector is useful for interpretations, it may sometimes not be a good predictor for real applications [25].

2.3.3 Interferents and drifts

Let n_k be the number of (known) absorbing species for which the corresponding concentrations are measured (called *analytes*), and n_u the number of (unknown) remaining absorbing species (called *interferents*). The concentration vector $\mathbf{z}$ can be partitioned into the n_k-dimensional known part $\mathbf{z}_k$ and the n_u-dimensional unknown part $\mathbf{z}_u$, with $n_s = n_k + n_u$. Similarly, $\mathbf{S}$, though unknown, can also be partitioned into the $[n_x \times n_k]$ pure-component spectra $\mathbf{S}_k$ of the n_k known analytes and the $[n_x \times n_u]$ pure-component spectra $\mathbf{S}_u$ of the n_u interferents:

$$\mathbf{z} = \begin{bmatrix}\mathbf{z}_k \\ \mathbf{z}_u\end{bmatrix}; \qquad \mathbf{S} = [\mathbf{S}_k \quad \mathbf{S}_u]. \tag{2.16}$$

Eq. (2.12) can thus be written as:

$$\tilde{\mathbf{x}} = \mathbf{S}_k\mathbf{z}_k + \mathbf{S}_u\mathbf{z}_u + \mathbf{v}_x. \tag{2.17}$$

The X-measurements may also be corrupted by systematic disturbances and offsets. Slowly-varying continuous systematic disturbances are caused by instrumental, operational and process changes, such as the effect of temperature, pressure and pH on the instrument, residue accumulation or aging of the instrument, and interactions between species [26–28]. Furthermore, infrequent discontinuous systematic disturbances in the X-measurements might also occur due to operational offsets, changes in probe alignment, addition of new species, cleaning and maintenance of instruments, differences in raw material quality used from batch to batch, or when data are collected on different instruments [29]. Let the term $\mathbf{d}_*$ correspond to an additive disturbance vector:

$$\tilde{\mathbf{x}} = \mathbf{S}_k\,\mathbf{z}_k + \mathbf{S}_u\mathbf{z}_u + \mathbf{d}_* + \mathbf{v}_x\,. \tag{2.18}$$

If $\mathbf{d}_*$ lies in a low-dimensional space of rank r_*, it can be interpreted as a contribution from r_* *pseudo-interferents*. Hence, for simplicity of notation, let the term $\mathbf{d}$ (henceforth referred to as *drift*) be the sum of the spectrum of the interferents and the disturbance vector:

$$\tilde{\mathbf{x}} = \mathbf{S}_k\,\mathbf{z}_k + \mathbf{d} + \mathbf{v}_x\,. \tag{2.19}$$

For n_c calibration measurements, Eq. (2.19) can be written in matrix form as:

$$\tilde{\mathbf{X}}_c = \mathbf{Z}_{c,k}\mathbf{S}_k^{\mathrm{T}} + \mathbf{D}_c + \mathbf{V}_{x,c}\,, \tag{2.20}$$

where $\mathbf{Z}_{c,k}$ is the known $[n_c \times n_k]$ concentration matrix of analytes, and $\mathbf{D}_c$ is the unknown $[n_c \times n_x]$ drift matrix.

2.3.4 Four sources of measured prediction error

Let a new spectrum be available according to Eq. (2.19), $\tilde{\mathbf{x}}_p = \mathbf{S}_k\,\mathbf{z}_{p,k} + \mathbf{d}_p + \mathbf{v}_{x,p}$. It is well-known that inverse calibration models of first-order instruments implicitly compensate for interferents that were present in the calibration data (see Appendix A). This motivates us to decompose the drift term $\mathbf{d}_p$ into a sum of two terms $\mathbf{d}_{p||c}$ and $\mathbf{d}_{p\perp c}$:

$$\mathbf{d}_p = \mathbf{d}_{p||c} + \mathbf{d}_{p\perp c}\,, \tag{2.21}$$

where, $\mathbf{d}_{p||c}$ lies in the drift subspace $\mathcal{R}(\mathbf{D}_c)$ *seen* in the calibration data, i.e. $\mathbf{d}_{p||c} = \mathbf{D}_c^{+}\mathbf{D}_c\mathbf{d}_p$, while $\mathbf{d}_{p\perp c}$ lies in the *unseen* drift subspace, i.e. $\mathbf{d}_{p\perp c} = (\mathbf{I}_{n_x} - \mathbf{D}_c^{+}\mathbf{D}_c)\mathbf{d}_p$. Using $\hat{y}_p = \tilde{\mathbf{x}}_p^{\mathrm{T}}\hat{\mathbf{b}}$, the measured prediction error $\tilde{y}_p - \hat{y}_p$ can be decomposed into four error terms:

$$\begin{aligned}\tilde{y}_p - \hat{y}_p &= y_p + v_{y,p} - \tilde{\mathbf{x}}_p^{\mathrm{T}}\hat{\mathbf{b}} \\ &= \underbrace{(y_p - \mathbf{z}_{p,k}^{\mathrm{T}}\,\mathbf{S}_k^{\mathrm{T}}\hat{\mathbf{b}} - \mathbf{d}_{p||c}^{\mathrm{T}}\hat{\mathbf{b}})}_{e1} - \underbrace{\mathbf{d}_{p\perp c}^{\mathrm{T}}\hat{\mathbf{b}}}_{e2} - \underbrace{\mathbf{v}_{x,p}^{\mathrm{T}}\,\hat{\mathbf{b}}}_{e3} + \underbrace{v_{y,p}}_{e4}\,, \end{aligned} \tag{2.22}$$

where

$e1$ is the error resulting from an inaccurate modeling of the latent subspace,

$e2$ is the error due to drift in prediction data,

$e3$ is the error due to X-noise in prediction data, and

$e4$ is the error due to y-noise in prediction data.

The error term $e1$, denoted as 'subspace modeling error' in this work, is due to X-noise and y-noise in the calibration data and the bias stemming from truncation in PCR/PLSR. Note that the four sources of error are independent of each other in the expectation sense.

For n_p prediction measurements, Eq. (2.19) can be written in matrix form as:

$$\tilde{\mathbf{X}}_p = \mathbf{Z}_{p,k}\mathbf{S}_k^{\mathrm{T}} + \mathbf{D}_p + \mathbf{V}_{x,p}\,. \tag{2.23}$$

Predictive ability is evaluated using the relative root-mean-squared error of prediction[2] (RRMSEP), defined as follows:

$$\mathrm{RRMSEP} := \frac{\sqrt{\sum_{j=1}^{j=n_p}(y_p(j) - \hat{y}_p(j))^2/n_p}}{R}\,, \tag{2.24}$$

where R is the response range. Since y_p may only be available with measurement noise, the *apparent* RRMSEP, abbreviated RRMSEP^{app}, is defined as follows [30]:

$$\mathrm{RRMSEP}^{app} := \frac{\sqrt{\sum_{j=1}^{j=n_p}(\tilde{y}_p(j) - \hat{y}_p(j))^2/n_p}}{R}\,. \tag{2.25}$$

[2] RRMSEP is an empirical estimate of the statistical risk

Let the RRMSEP be defined separately for each error term:

$$\begin{aligned}
\mathrm{RRMSEP}_{e1} = \mathrm{RRMSEP}_{e1}^{app} &= \frac{\sqrt{\sum_{j=1}^{j=n_p}(y_p(j) - \mathbf{z}_{p,k}^{\mathrm{T}}(j)\,\mathbf{S}_k^{\mathrm{T}}\hat{\mathbf{b}} - \mathbf{d}_{p||c}^{\mathrm{T}}\hat{\mathbf{b}})^2/n_p}}{R} \\
\mathrm{RRMSEP}_{e2} = \mathrm{RRMSEP}_{e2}^{app} &= \frac{\sqrt{\sum_{j=1}^{j=n_p}(\mathbf{d}_{p\perp c}^{\mathrm{T}}(j)\hat{\mathbf{b}})^2/n_p}}{R} \\
\mathrm{RRMSEP}_{e3} = \mathrm{RRMSEP}_{e3}^{app} &= \frac{\sqrt{\sum_{j=1}^{j=n_p}(\mathbf{v}_{x,p}^{\mathrm{T}}(j)\,\hat{\mathbf{b}})^2/n_p}}{R} \\
\mathrm{RRMSEP}_{e4}^{app} &= \frac{\sqrt{\sum_{j=1}^{j=n_p}(v_{y,p}(j))^2/n_p}}{R}\,.
\end{aligned} \tag{2.26}$$

Since the error term $e4$ stems from the limitation of the reference instrument used for validation purposes, typically one does not have a handle on $\mathrm{RRMSEP}_{e4}^{app}$. $\mathrm{RRMSEP}_{e3}^{app}$ can be kept low either by using replicate measurements to average out the measurement noise, or by regularizing $\hat{\mathbf{b}}$ such that $||\hat{\mathbf{b}}||$ is reduced (e.g. by choosing fewer factors in PCR/PLSR). Chapter 3 studies drift subspace correction to reduce RRMSEP_{e2}, Chapter 4 studies latent subspace correction to reduce RRMSEP_{e1}, and Chapter 5 studies data reconciliation to reduce the overall RRMSEP.

3

Drift subspace correction using master/slave data

This chapter studies drift subspace correction (to reduce RRMSEP_{e2}) based on additional information in the form of master/slave data.

3.1 Overview

Let the calibration and prediction data be available with drift according to Eqs. (2.20) and (2.23). Drift can make the calibration model unsuitable for prediction and necessitates re-calibration. This chapter studies measurement-based drift correction for inverse calibration. A schematic diagram of two approaches for re-calibration is shown in Fig. 3.1. The first solution (see Fig. 3.1a) is to repeat the whole calibration procedure and rebuild a new model for the new instrumental, operational and process conditions. To obtain similar prediction accuracy as with the old calibration model, a similar large number of data are required for the full re-calibration. This is costly because it requires reference concentrations for many samples. The second solution (see Fig. 3.1b) exploits the fact that spectroscopic data typically retain structural similarities under different conditions. These similarities can be exploited by reusing the old calibration data together with *master/slave data* obtained under new conditions. The amount of master/slave data required is typically less than that for full re-calibration, thereby circumventing the need for many new reference concentrations. Drift correction that proceeds by adding master/slave data with drift to the calibration data, is referred to as *implicit correction method* (ICM),

while an *explicit correction method* (ECM) estimates the drift subspace based on calibration and master/slave data and makes the new calibration model less sensitive to drift. ICM and ECM are shown to be equivalent in the absence of noise. However, they differ in the way they handle noise.

For inverse calibration problems, several ECMs exist in the literature, e.g. component correction (CC) [31], independent interference reduction (IIR) [32], generalized least squares (GLS) [33, 34], external parameter orthogonalization (EPO) [35], calibration transfer by orthogonal projection (TOP) [36], dynamic orthogonal projection (DOP) [37, 38], difference correction of prediction spectra (DCPS) [39], and error removal by orthogonal subtraction (EROS) [40]. Confronted with a choice from this alphabet soup of ECMs only adds to the confusion of a practitioner. It is the endeavor of this chapter to show that these ECMs can be cast in a two-step framework. In the first step, the drift subspace is estimated using different types of master/slave data. In the second step, the calibration data lying in the estimated drift subspace is scaled down either partially using *shrinkage* or fully using *orthogonal projection* (OP).

The different ECMs for inverse calibration are investigated analytically and with experimental data. The first example studies the validity of a key assumption, that typically *drift lies in a low-dimensional space*. The next three examples study drift correction on one instrument (temperature effects, spectral differences between samples obtained from different plants, instrumental drift), while the fifth example studies calibration transfer between two instruments. The chapter is organized as follows. Sections 3.2 and 3.3 propose the two-step framework for several ECMs proposed in the literature and discusses their essential differences. Section 3.4 discusses two propositions on the equivalence of various methods. Five experimental studies then illustrate drift correction in Section 3.5. Related methods are discussed in Section 3.6, and Section 3.7 concludes the chapter.

3.2 Drift subspace estimation (Step 1)

ECMs involve the two steps of drift-space estimation and drift correction. The first step is described in this section.

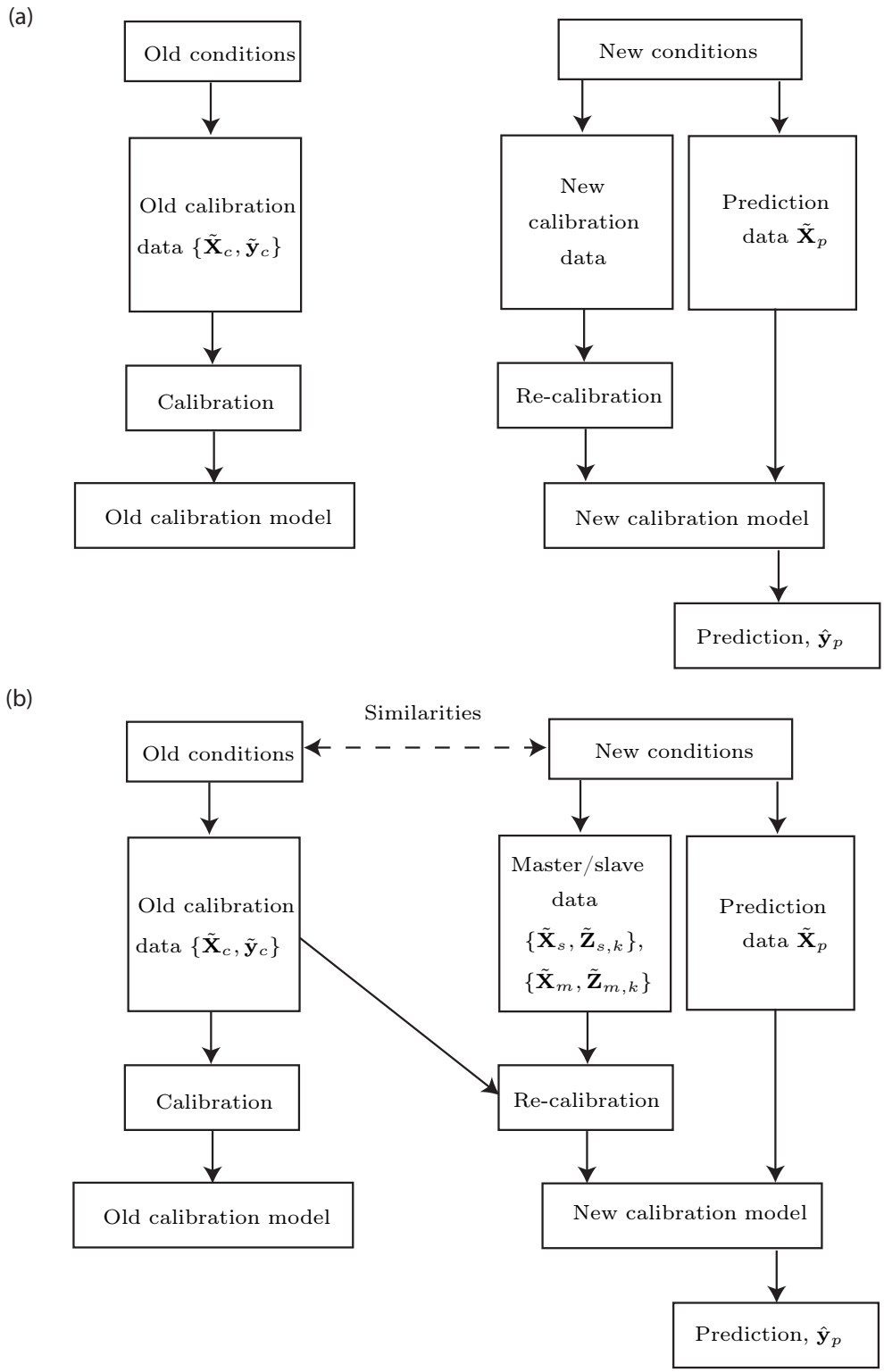


Fig. 3.1. (a) Full re-calibration for new conditions, (b) drift correction using master/slave data.

3.2.1 Master/slave data

The drift subspace is estimated from n_d master/slave samples that are available on-line or off-line in addition to the calibration data. The master/slave data, which encompass the four matrices $\tilde{\mathbf{X}}_s$, $\tilde{\mathbf{Z}}_{s,k}$, $\tilde{\mathbf{X}}_m$ and $\tilde{\mathbf{Z}}_{m,k}$, are of two types as indicated next:

(i) the measured $[n_d \times n_x]$ *slave spectroscopic matrix* $\tilde{\mathbf{X}}_s$ with unseen drift:

$$\tilde{\mathbf{X}}_s = \mathbf{Z}_{s,k}\mathbf{S}_k^{\mathrm{T}} + \mathbf{D}_s + \mathbf{V}_{x,s}, \tag{3.1}$$

(ii) the $[n_d \times n_k]$ *slave concentration matrix* $\tilde{\mathbf{Z}}_{s,k}$ of the n_k analytes,

(iii) the $[n_d \times n_x]$ *master spectroscopic matrix* $\tilde{\mathbf{X}}_m$ (without unseen drift) computed as a linear combination of $\tilde{\mathbf{X}}_s$ or $\tilde{\mathbf{X}}_c$ using a $[n_d \times n_d]$ linear operator $\mathbf{A}$:

$$\begin{aligned} \tilde{\mathbf{X}}_m &= \mathbf{A}\tilde{\mathbf{X}}_s \qquad \text{(Type 1)} \\ \tilde{\mathbf{X}}_m &= \mathbf{A}\tilde{\mathbf{X}}_c \qquad \text{(Type 2).} \end{aligned} \tag{3.2}$$

If $\tilde{\mathbf{X}}_m$ contains only seen drift, then it can alternatively be written as:

$$\tilde{\mathbf{X}}_m = \mathbf{Z}_{m,k}\mathbf{S}_k^{\mathrm{T}} + \mathbf{D}_m + \mathbf{V}_{x,m}, \tag{3.3}$$

with $\mathbf{D}_m$ such that $\mathcal{R}(\mathbf{D}_m) \subseteq \mathcal{R}(\mathbf{D}_c)$, and

(iv) the $[n_d \times n_k]$ *master concentration matrix* $\tilde{\mathbf{Z}}_{m,k}$ of the n_k analytes with the property following from Eq. (3.2):

$$\begin{aligned} \tilde{\mathbf{Z}}_{m,k} &= \mathbf{A}\tilde{\mathbf{Z}}_{s,k} \qquad \text{(Type 1)} \\ \tilde{\mathbf{Z}}_{m,k} &= \mathbf{A}\tilde{\mathbf{Z}}_{c,k} \qquad \text{(Type 2).} \end{aligned} \tag{3.4}$$

Since the drift affects the spectroscopic measurements but not the concentrations, the following condition must be satisfied:

$$\tilde{\mathbf{Z}}_{m,k} = \tilde{\mathbf{Z}}_{s,k}. \tag{3.5}$$

Note that the concentrations of the interferents, $\mathbf{Z}_{s,u}$, can take any arbitrary values. A suitable linear operator $\mathbf{A}$ is such that Eqs. (3.2), (3.3), (3.4), and

(3.5) are simultaneously satisfied. In Type 1 data, $\mathbf{A}$ is deduced from the knowledge of $\mathbf{X}_m$, while in Type 2 data, $\mathbf{A}$ is computed from the knowledge of $\mathbf{Z}_{m,k}$. Some concrete examples of $\mathbf{A}$ will be discussed in Sections 3.2.2 and 3.2.3.

The basic idea is to estimate the drift subspace from the difference between $\tilde{\mathbf{X}}_s$ and $\tilde{\mathbf{X}}_m$:

$$\hat{\mathbf{D}} = \tilde{\mathbf{X}}_s - \tilde{\mathbf{X}}_m \,. \tag{3.6}$$

3.2.2 Type 1: Operator A deduced from knowledge of $\mathbf{X}_m$

In [35,39,40], master/slave data at different temperatures are used to make the calibration model robust against temperature effects. In general, if q samples are measured at t different conditions (e.g. temperatures, pH, particle forms and sizes, sample preparation and suppliers), the mean spectrum (in fact, any linear combination) of the t slave spectra can be considered as the master spectrum for the jth sample. Let $\mathbf{1}_t$ be a t-dimensional vector of ones, and the *a priori* known operator $\mathbf{A}$ correspond to mean-centering, i.e. $\mathbf{A} = \frac{1}{t}\mathbf{1}_t\mathbf{1}_t^{\mathrm{T}}$. For the q samples, $n_d = q\,t$ and:

$$\tilde{\mathbf{X}}_s = \begin{bmatrix} \mathbf{x}_{s,1} \\ \vdots \\ \mathbf{x}_{s,q} \end{bmatrix}; \qquad \tilde{\mathbf{X}}_m = \begin{bmatrix} \mathbf{x}_{m,1} \\ \vdots \\ \mathbf{x}_{m,q} \end{bmatrix} \qquad \text{with} \tag{3.7}$$

$$\mathbf{X}_{s,j} = \begin{bmatrix} \mathbf{x}_{j,1}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{j,t}^{\mathrm{T}} \end{bmatrix}; \qquad \mathbf{X}_{m,j} = \mathbf{A}\,\mathbf{X}_{s,j} \qquad \forall j = 1, \ldots, q \,. \tag{3.8}$$

Since the analyte concentrations of the slave and master are identical (but possibly unknown) for each of the q samples, it can be verified that $\tilde{\mathbf{Z}}_{m,k,j} = \mathbf{A}\tilde{\mathbf{Z}}_{s,k,j} = \tilde{\mathbf{Z}}_{s,k,j}$, $\forall j = 1, \ldots, q$. Note that the concentrations need not be known explicitly for drift-space estimation. Furthermore, mean-centering in Eq. (3.8) can potentially cause rank deficiency, thereby removing one factor of the drift subspace for each sample. For example, if the drift is a sample-invariant constant baseline, this drift is not captured in $\hat{\mathbf{D}}$.

In [33, 34, 36], q samples are used for calibration transfer between $t = 2$ instruments. Here, a reasonable choice of $\tilde{\mathbf{X}}_m$ corresponds to the spectroscopic measurements on the instrument corresponding to the calibrated instrument

(say, first instrument). With $\tilde{\mathbf{X}}_s$ and $\tilde{\mathbf{X}}_m$ defined as:

$$\tilde{\mathbf{X}}_s = \begin{bmatrix} \mathbf{x}_{s,1} \\ \mathbf{x}_{s,2} \end{bmatrix}; \qquad \tilde{\mathbf{X}}_m = \mathbf{A}\,\tilde{\mathbf{X}}_s = \begin{bmatrix} \mathbf{x}_{s,1} \\ \mathbf{x}_{s,1} \end{bmatrix}, \tag{3.9}$$

this corresponds to $\mathbf{A} = \begin{bmatrix} \mathbf{I}_q & \mathbf{0}_{[q\times q]} \\ \mathbf{I}_q & \mathbf{0}_{[q\times q]} \end{bmatrix}$. It is easy to verify that $\mathbf{A}$ satisfies $\tilde{\mathbf{Z}}_{m,k} = \mathbf{A}\tilde{\mathbf{Z}}_{s,k} = \tilde{\mathbf{Z}}_{s,k}$.

In [32, 41], $\tilde{\mathbf{X}}_s$ are measured for samples containing negligible amounts of the analytes of interest, i.e. $\tilde{\mathbf{Z}}_{s,k} = \mathbf{0}$. Here, since $\tilde{\mathbf{X}}_s$ is due to factors that are unrelated to the analytes of interest, a reasonable choice for the master spectroscopic matrix is $\tilde{\mathbf{X}}_m = \mathbf{0}$. This corresponds to the operator $\mathbf{A} = \mathbf{0}$, which satisfies $\tilde{\mathbf{Z}}_{m,k} = \mathbf{0} = \tilde{\mathbf{Z}}_{s,k}$.

In [31], $\tilde{\mathbf{X}}_s$ are known to contain significant analyte information, however, the analyte is held constant. Therefore, the estimated drift is mean-centered to remove all analyte information. This way, it becomes a special case of Eqs. (3.7) and (3.8) with $q = 1$.

3.2.3 Type 2: Operator A computed from knowledge of $\mathbf{Z}_{m,k}$

In [37], an alternate way for capturing the drift subspace is proposed. It uses n_d on-line measurements $\tilde{\mathbf{X}}_s$ together with the corresponding concentrations $\tilde{\mathbf{Z}}_{s,k}$ that are either measured by reference analytics or known *a priori* (e.g. known initial concentrations in a batch reactor). With on-line measurements, $\tilde{\mathbf{X}}_m$ is not known beforehand, and must be computed using known values of $\tilde{\mathbf{Z}}_{m,k} = \tilde{\mathbf{Z}}_{s,k}$ and the calibration data. Here, the $[n_d \times n_c]$ matrix $\mathbf{A}$ is based on a linear combination of the calibration data such that $\tilde{\mathbf{Z}}_{m,k} = \tilde{\mathbf{Z}}_{s,k} = \mathbf{A}\,\tilde{\mathbf{Z}}_{c,k}$. Hence, $\mathbf{A}$ satisfies the condition:

$$\tilde{\mathbf{Z}}_{s,k} = \mathbf{A}\,\tilde{\mathbf{Z}}_{c,k}\,, \tag{3.10}$$

leading to

$$\tilde{\mathbf{X}}_m = \mathbf{A}\,\tilde{\mathbf{X}}_c\,. \tag{3.11}$$

For Type 2, Eqs. (3.2) and (3.3) are simultaneously satisfied by definition. However, if a solution to Eq. (3.10) does not exist, Eqs. (3.4) and (3.5) are not simul-

taneously satisfied. Any $\mathbf{A}$ satisfying Eq. (3.10) leads to an unbiased estimate of the drift subspace, e.g. a minimum-norm solution $\mathbf{A} = \tilde{\mathbf{Z}}_{s,k}\tilde{\mathbf{Z}}_{c,k}^{+}$.

In [38], an injection method was proposed that involves adding known amounts of analytes $\tilde{\mathbf{Z}}_{s,k} = \tilde{\mathbf{Z}}_{m,k}$. $\tilde{\mathbf{X}}_s$ is the corresponding difference spectrum before and after the addition, while $\tilde{\mathbf{X}}_m$ is the expected difference spectrum computed as in Eq. (3.11). However, the injection method does not always capture the drift. For example, if the drift spectrum before and after the addition remains the same, the difference spectrum does not contain any information regarding this drift.

3.3 Drift correction (Step 2)

In Step 2, the calibration data lying in the estimated drift subspace are scaled down using shrinkage or orthogonal projection (OP), as discussed in the following.

3.3.1 Shrinkage

In GLS [33,34], $\hat{\mathbf{b}}$ is computed from the calibration data pair $\{\tilde{\mathbf{X}}_c\mathbf{H}, \tilde{\mathbf{y}}_c\}$, where the $[n_x \times n_x]$ weighting matrix $\mathbf{H}$ is used to downweigh (or shrink) the drift subspace:

$$\mathbf{H} = \left(\frac{1}{n_d - 1}\hat{\mathbf{D}}^{\mathrm{T}}\hat{\mathbf{D}} + \alpha^2\mathbf{I}_{n_x}\right)^{-\frac{1}{2}} . \tag{3.12}$$

The positive real parameter α determines the extent of shrinkage; as α decreases, the drift subspace is shrunk more. Typically, α is adjusted for each analyte. For a prediction spectrum $\tilde{\mathbf{x}}_p$, $\hat{y}_p = \tilde{\mathbf{x}}_p^{\mathrm{T}}\hat{\mathbf{b}}^*$, where $\hat{\mathbf{b}}^* = \mathbf{H}\,\hat{\mathbf{b}}$. Note that $\hat{\mathbf{b}}^*$ is invariant to multiplication of $\mathbf{H}$ by any scalar.

3.3.2 Orthogonal projection (OP)

In CC [31], EPO [35], TOP [36], DOP [37,38], DCPS [39], and EROS [40], only r_d significant loadings of $\hat{\mathbf{D}}$, computed by PCA, are retained:

$$\begin{aligned} \hat{\mathbf{D}} &= \mathbf{T}\,\mathbf{P}^{\mathrm{T}} + \mathbf{E} \\ \hat{\hat{\mathbf{D}}} &= \mathbf{T}\,\mathbf{P}^{\mathrm{T}}, \end{aligned} \tag{3.13}$$

where $\mathbf{T}$ and $\mathbf{P}$ are the $[n_d \times r_d]$ scores matrix and the $[n_x \times r_d]$ loading matrix, respectively. The $[n_x \times n_x]$ orthogonal projection matrix $\hat{\mathbf{N}} = (\mathbf{I}_{n_x} - \mathbf{P}\mathbf{P}^{\mathrm{T}})$ is computed, and $\hat{\mathbf{b}}$ is obtained from calibration with the data pair $\{\tilde{\mathbf{X}}_c\hat{\mathbf{N}}, \tilde{\mathbf{y}}_c\}$. Hence, $\hat{\mathbf{b}} \in \hat{\mathbf{N}}$ and $\hat{\mathbf{b}}^* = \hat{\mathbf{N}}\hat{\mathbf{b}} = \hat{\mathbf{b}}$. For a prediction sample $\tilde{\mathbf{x}}_p$, $\hat{y}_p = (\tilde{\mathbf{x}}_p^{\mathrm{T}}\hat{\mathbf{N}})\,\hat{\mathbf{b}} = \tilde{\mathbf{x}}_p^{\mathrm{T}}\hat{\mathbf{b}}$.

3.3.3 Choice of the meta-parameters r_d and α

Specifying the meta-parameters r_d for OP and α for shrinkage is a delicate task. Ideally, the choice of these meta-parameters should be based on the four sources of prediction error (see Eq. (2.22)) analyzed simultaneously. The equation is repeated here with the recalibrated $\hat{\mathbf{b}}^*$:

$$\tilde{y}_p - \hat{y}_p = \underbrace{(y_p - \mathbf{z}_{p,k}^{\mathrm{T}}\,\mathbf{S}_k^{\mathrm{T}}\hat{\mathbf{b}}^* - \mathbf{d}_{p||c}^{\mathrm{T}}\hat{\mathbf{b}}^*)}_{e1} - \underbrace{\mathbf{d}_{p\perp c}^{\mathrm{T}}\hat{\mathbf{b}}^*}_{e2} - \underbrace{\mathbf{v}_{x,p}^{\mathrm{T}}\,\hat{\mathbf{b}}^*}_{e3} + \underbrace{v_{y,p}}_{e4}\,. \tag{3.14}$$

Let $\mathbf{s}_i$ be the pure-component spectrum of the i^{th} analyte, and the overlap of $\mathbf{d}_p$ with $\mathbf{s}_i$ be defined by a parameter $\gamma = \frac{|\mathbf{s}_i^{\mathrm{T}}\hat{\mathbf{b}}^*|}{||\mathbf{s}_i||\,||\hat{\mathbf{b}}^*||}$, γ varying between 0 (full drift overlap) and 1 (no overlap). An ideal $\hat{\mathbf{b}}^*$ would be such that error terms $e1 = e2 = e3 = 0$, which means that (i) $\hat{\mathbf{b}}^*$ is orthogonal to $\mathbf{d}_{p||c}$, $\mathbf{d}_{p\perp c}$, and the pure-component spectra excluding the analyte of interest, (ii) $\mathbf{s}_i^{\mathrm{T}}\hat{\mathbf{b}}^* = 1$, and (iii) $||\hat{\mathbf{b}}^*||$ tends to zero. However, satisfying these three conditions simultaneously is impossible for the following reasons: Firstly, because of the noise in the calibration and master/slave data, Conditions (i) and (ii) can only be approximately satisfied. Secondly, Conditions (ii) and (iii) can obviously not be satisfied simultaneously. This results in a trade-off between the first three sources of prediction error in Eq. (3.14).

Shrinkage or OP reduce the effect of error term $e2$ but increase the effects of error terms $e1$ and $e3$. Error term $e1$ increases if drift overlaps significantly with $\mathbf{s}_i$, thereby reducing the NAS component and the signal-to-noise ratio (SNR). Also, error term $e3$ increases because reduction in the NAS component results in an increase in $||\hat{\mathbf{b}}^*||$.

The trade-off is regulated in OP by the parameter r_d (which determines how many significant drift directions are considered in the estimated drift subspace), and in shrinkage by the parameter α (which determines how aggressively the estimated drift subspace is shrunk). If a large number of samples (n_d) were available for drift correction, a separate cross-validation could be performed to determine the meta-parameters. However, a large n_d defeats the purpose of migrating the calibration model to the new conditions with unseen drift using only a few samples. For small n_d and in the absence of prior information about the pure-component spectra, the drift component and the noise variance, the choice of the meta-parameters is often guided by rules of thumb or heuristic criteria based on expert opinion.

3.4 Equivalence of methods

3.4.1 Proposition 1: Equivalence of shrinkage and OP

The following proposition specifies the conditions for which OP and no correction become special cases of shrinkage. Let σ_1 and $\sigma_{r_{max}}$ be the maximum and minimum singular values of $\hat{\mathbf{D}}$, respectively, with $r_{max} = \text{rank}\left(\hat{\mathbf{D}}\right)$.

Proposition 1 *Drift correction by OP is a special case of shrinkage for $r_d = r_{max}$ and $\alpha/\sigma_{r_{max}} \to 0$. For $\alpha/\sigma_1 \to \infty$, no correction is performed.*

Proof:

Let the SVD of $\frac{1}{n_d-1}\hat{\mathbf{D}}^{\mathrm{T}}\hat{\mathbf{D}} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, where the dimensions of $\mathbf{V}$ and $\mathbf{\Sigma}$ are both $[n_x \times n_x]$. Only the first r_{max} diagonal elements of $\mathbf{\Sigma}$ are nonzero, i.e. $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_{r_{max}}, 0, \ldots, 0)$. Since $\mathbf{V}\mathbf{V}^{\mathrm{T}}$ is a projection matrix spanning the complete n_x-dimensional space, $\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{I}_{n_x}$. From Eq. (3.12):

$$\begin{aligned}
\mathbf{H} &= \left(\frac{1}{n_d-1}\hat{\mathbf{D}}^{\mathrm{T}}\hat{\mathbf{D}} + \alpha^2\mathbf{I}_{n_x}\right)^{-\frac{1}{2}} \\
&= (\mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} + \mathbf{V}\alpha^2\mathbf{V}^{\mathrm{T}})^{-\frac{1}{2}} \\
&= (\mathbf{V}(\mathbf{\Sigma} + \alpha^2\mathbf{I}_{n_x})\mathbf{V}^{\mathrm{T}})^{-\frac{1}{2}} \\
&= \mathbf{V}(\mathbf{\Sigma} + \alpha^2\mathbf{I}_{n_x})^{-\frac{1}{2}}\mathbf{V}^{\mathrm{T}}\,. \qquad (3.15)
\end{aligned}$$

Thus, shrinkage with $\mathbf{H}$ results in a projection of rows of $\tilde{\mathbf{X}}_c$ in the n_x-dimensional $\mathcal{R}(\mathbf{V}^{\mathrm{T}})$. The resulting scores $\tilde{\mathbf{X}}_c\mathbf{V}$ are shrunk by the corresponding shrinkage coefficients $\left\{\frac{1}{\sqrt{\sigma_1+\alpha^2}},\ldots,\ \frac{1}{\sqrt{\sigma_{r_{max}}+\alpha^2}},\frac{1}{\alpha},\ldots,\frac{1}{\alpha}\right\}$. Since $\hat{\mathbf{b}}^*$ is invariant to multiplication of $\mathbf{H}$ by any scalar, the shrinkage coefficients can also be expressed as $\left\{\frac{\alpha}{\sqrt{\sigma_1+\alpha^2}},\ldots,\frac{\alpha}{\sqrt{\sigma_{r_{max}}+\alpha^2}},1,\ldots,1\right\}$. Hence, as $\alpha/\sigma_{r_{max}}\to 0$, the first r_{max} coefficients corresponding to drift directions tend to 0, implying that shrinkage tends to OP. Furthermore, if $\alpha/\sigma_1\to\infty$, the shrinkage coefficients are $\{1,\ldots,1\}$, implying that no correction is performed. □

It can be observed from Eq. (3.15) that each drift direction (column of $\mathbf{V}$) is scaled down as the square root of the sum of its eigenvalue and α^2.

3.4.2 Proposition 2: Equivalence of ICM and ECM with OP

Consider the following assumptions:

A1: Eqs. (2.20), (2.23), (3.1), and (3.3) are valid and the noise is negligible,

A2: a suitable $\mathbf{A}$ exists, i.e. Eqs. (3.2), (3.3), (3.4), and (3.5) are simultaneously satisfied,

A3: $\operatorname{rank}\left(\begin{bmatrix}\mathbf{S}_k^{\mathrm{T}}\\ \mathbf{D}_c\\ \mathbf{D}_s\end{bmatrix}\right)=n_k+\operatorname{rank}\left(\begin{bmatrix}\mathbf{D}_c\\ \mathbf{D}_s\end{bmatrix}\right)$, i.e. the seen drift space does not fully overlap with the signal space, and $\operatorname{rank}\left(\begin{bmatrix}\mathbf{Z}_{c,k}\\ \mathbf{Z}_{s,k}\end{bmatrix}\right)=n_k$, i.e. the matrix containing the calibration and slave concentrations of the known analytes is full rank, and

A4: $\mathcal{R}(\mathbf{D}_p)\subseteq\mathcal{R}(\begin{bmatrix}\mathbf{D}_s\\ \mathbf{D}_c\end{bmatrix})$, i.e. the unseen drift space in the prediction spectroscopic data lies within the union of the drift spaces seen in the calibration and slave data.

Proposition 2 *Let the working assumptions A1–A4 hold and $\mathcal{R}(\hat{\mathbf{D}})$ be estimated as:*

$$\mathcal{R}(\hat{\mathbf{D}})=\mathcal{R}(\tilde{\mathbf{X}}_s-\tilde{\mathbf{X}}_m)\,. \tag{3.16}$$

Then, the $[n_x\times n_k]$ regression matrix $\hat{\mathbf{B}}_k^{\mathrm{aug}}$ from ICM,

$$\hat{\mathbf{B}}_k^{\mathrm{aug}}=\begin{bmatrix}\mathbf{X}_c\\ \mathbf{X}_s\end{bmatrix}^{+}\begin{bmatrix}\mathbf{Z}_{c,k}\\ \mathbf{Z}_{s,k}\end{bmatrix}\,, \tag{3.17}$$

and the $[n_x \times n_k]$ *regression matrix* $\hat{\mathbf{B}}_k^{\text{proj}}$ *from ECM with OP,*

$$\hat{\mathbf{B}}_k^{\text{proj}} = \left(\mathbf{X}_c(\mathbf{I}_{n_x} - \hat{\mathbf{D}}^+\hat{\mathbf{D}})\right)^+ \mathbf{Z}_{c,k}\,, \tag{3.18}$$

are identical and, furthermore:

$$\mathbf{D}_p\hat{\mathbf{B}}_k^{\text{proj}} = \mathbf{D}_p\hat{\mathbf{B}}_k^{\text{aug}} = \mathbf{0}_{[n_p \times n_k]}\,. \tag{3.19}$$

Proof:

Without loss of generality, consider calibration for the i^{th} analyte, and let $\mathbf{S}_{k,-i}$ contain all but the i^{th} column of $\mathbf{S}_k$. With Assumption A3, calibration with noise-free data leads to $\hat{\mathbf{b}} \propto$ pNAS vector (see Section 2.3.2). The pNAS vector for $\left\{\begin{bmatrix}\mathbf{X}_c\\ \mathbf{X}_s\end{bmatrix}, \begin{bmatrix}\mathbf{y}_c\\ \mathbf{y}_s\end{bmatrix}\right\}$ is the projection of $\mathbf{s}_i$ on the space orthogonal to $\mathcal{R}(\mathbf{S}_{k,-i}^{\mathrm{T}}) \cup \mathcal{R}(\mathbf{D}_c) \cup \mathcal{R}(\mathbf{D}_s)$. Similarly, the pNAS vector for $\left\{\left(\mathbf{X}_c(\mathbf{I}_{n_x} - \hat{\mathbf{D}}^+\hat{\mathbf{D}})\right), \mathbf{y}_c\right\}$ is the projection of $(\mathbf{I}_{n_x} - \hat{\mathbf{D}}^+\hat{\mathbf{D}})\mathbf{s}_i$ on the space orthogonal to $\mathcal{R}\left(\mathbf{S}_{k,-i}^{\mathrm{T}}(\mathbf{I}_{n_x} - \hat{\mathbf{D}}^+\hat{\mathbf{D}})\right) \cup \mathcal{R}\left(\mathbf{D}_c(\mathbf{I}_{n_x} - \hat{\mathbf{D}}^+\hat{\mathbf{D}})\right)$. This is equivalent to the projection of $\mathbf{s}_i$ on the space orthogonal to $\mathcal{R}(\mathbf{S}_{k,-i}^{\mathrm{T}}) \cup \mathcal{R}(\mathbf{D}_c) \cup \mathcal{R}(\hat{\mathbf{D}})$.

Using $\mathbf{Z}_{s,k} - \mathbf{Z}_{m,k} = \mathbf{0}$ (which follows from Assumption A2),

$$\begin{aligned}\mathcal{R}(\hat{\mathbf{D}}) &= \mathcal{R}(\mathbf{X}_s - \mathbf{X}_m)\\ &= \mathcal{R}\left(\mathbf{D}_s - \mathbf{D}_m + (\mathbf{Z}_{s,k} - \mathbf{Z}_{m,k})\mathbf{S}_k^{\mathrm{T}}\right)\\ &= \mathcal{R}(\mathbf{D}_s - \mathbf{D}_m)\,.\end{aligned} \tag{3.20}$$

Multiplying both sides by $(\mathbf{I}_{n_x} - \mathbf{D}_c^+\mathbf{D}_c)$, and using $\mathcal{R}(\mathbf{D}_m) \subseteq \mathcal{R}(\mathbf{D}_c)$, gives:

$$\begin{aligned}\mathcal{R}\left(\hat{\mathbf{D}}(\mathbf{I}_{n_x} - \mathbf{D}_c^+\mathbf{D}_c)\right) &= \mathcal{R}\left((\mathbf{D}_s - \mathbf{D}_m)\,(\mathbf{I}_{n_x} - \mathbf{D}_c^+\mathbf{D}_c)\right)\\ &= \mathcal{R}\left(\mathbf{D}_s\,(\mathbf{I}_{n_x} - \mathbf{D}_c^+\mathbf{D}_c)\right)\,.\end{aligned} \tag{3.21}$$

Since Eq. (3.21) says that $\mathcal{R}(\hat{\mathbf{D}})$ may differ from $\mathcal{R}(\mathbf{D}_s)$ only in the space of $\mathcal{R}(\mathbf{D}_c)$, one can write:

$$\mathcal{R}(\hat{\mathbf{D}}) \cup \mathcal{R}(\mathbf{D}_c) = \mathcal{R}(\mathbf{D}_s) \cup \mathcal{R}(\mathbf{D}_c)\,, \tag{3.22}$$

and

$$\mathcal{R}(\mathbf{S}_{k,-i}^{\mathrm{T}}) \cup \mathcal{R}(\mathbf{D}_c) \cup \mathcal{R}(\mathbf{D}_s) = \mathcal{R}(\mathbf{S}_{k,-i}^{\mathrm{T}}) \cup \mathcal{R}(\mathbf{D}_c) \cup \mathcal{R}(\hat{\mathbf{D}})\,. \tag{3.23}$$

Hence, the pNAS of the two calibrations are identical. Furthermore, since $\mathcal{R}(\mathbf{D}_p) \subseteq \mathcal{R}(\left[\begin{smallmatrix}\mathbf{D}_s\\ \mathbf{D}_c\end{smallmatrix}\right])$ (Assumption A4), $\mathbf{D}_p\hat{\mathbf{b}} = \mathbf{0}_{n_p}$. □

Proposition 2 says that, under Assumptions A1–A3, ICM and ECM with OP are equivalent. Furthermore, if Assumption A4 is also valid, ICM and ECM with OP lead to correct prediction. Note that Proposition 2 holds for the noise-free case. In the presence of noise, ICM and ECM with OP differ.

3.5 Illustrative examples

3.5.1 First example (experimental data): Drift lies in a low-dimensional space

This data set is from the "Shootout" at Chimiométrie 2007 involving wheat samples analyzed for protein content [42]. The data involves artificial variations created by varying the moisture and the particle size of wheat samples, measured at different temperatures, and on different instruments. The X-measurements are NIR transmittance spectra recorded at 550 wavelengths over the range of 1300-2398 nm on a Foss NIRSystems 4500 spectrometer, while the weight percentage of protein content is used as the response variable.

Three sets of data are available. The first set consists of 10 X-measurements on 31 instruments (Dataset 1). The second set consists of a different set of 10 X-measurements on 17 instruments (Dataset 2). Also, 11 whole grain samples are moistened at two levels, ground to two particle sizes, and two replicate X-measurements are obtained at three temperatures (Dataset 3). Hence, Dataset 1 contains $10 \times 31 = 310$ spectra, Dataset 2 contains $10 \times 17 = 170$ spectra, and Dataset 3 contains $11 \times 2 \times 2 \times 3 \times 2 = 264$ spectra. The goal of this study is to show that drift due to instrument, moisture, particle size, or temperature differences lies in a low-dimensional space. Since the reference protein values are unknown, master/slave data of Type 1 are used for drift-space estimation.

Datasets 1 and 2 are used to study instrument drift. For Dataset 1, $\tilde{\mathbf{X}}_s$ and $\tilde{\mathbf{X}}_m$ are obtained as discussed in Eqs. (3.7)–(3.8) using all the 10 samples, leading to $\hat{\mathbf{D}}$ of size $[310 \times 550]$.

Several statistical tests exist in the literature that help to determine the pseudo-rank of $\hat{\mathbf{D}}$, e.g. Wilks' λ test, Malinowski's F-test, and Faber-Kowalski F-test [43,44]. Based on the analysis of the eigenvalues of the covariance matrix $\hat{\mathbf{D}}^{\mathrm{T}}\hat{\mathbf{D}}$, these tests attempt to distinguish between the eigenvalues due to signal and noise. However, results based on random matrix theory and perturbation theory show that even for randomly generated matrices, few principal components may capture significant amount of variation in the data [45]. Hence, instead of trying to answer the question *"What is the pseudo-rank of* $\hat{\mathbf{D}}$*?"*, we answer a related but more pertinent question *"Let drift subspace be estimated from a few samples. Does drift in new samples lie in the same subspace?"*. A subspace comparison method is discussed in [46]. However, in accordance with the philosophy of cross-validation, we partition $\hat{\mathbf{D}}$ into the two blocks $\hat{\mathbf{D}}_1$ and $\hat{\mathbf{D}}_2$, where $\hat{\mathbf{D}}_1$ is used for model fitting, and $\hat{\mathbf{D}}_2$ for testing. Choosing two samples randomly out of the ten, $\hat{\mathbf{D}}$ can be partitioned as:

$$\hat{\mathbf{D}} = \begin{bmatrix} \hat{\mathbf{D}}_1 \\ \hat{\mathbf{D}}_2 \end{bmatrix} , \tag{3.24}$$

where $\hat{\mathbf{D}}_1$ of dimension $[62 \times 550]$ consists of 31 measurements each from the chosen two samples, and $\hat{\mathbf{D}}_2$ of dimension $[248 \times 550]$ consists of 31 measurements each from the remaining eight samples. SVD of $\hat{\mathbf{D}}_1$ and $\hat{\mathbf{D}}_2$ leads to:

$$\begin{aligned} \hat{\mathbf{D}}_1 = &\quad \mathbf{U}_1 \, \mathbf{\Sigma}_1 \, \mathbf{V}_1^{\mathrm{T}} \\ \hat{\mathbf{D}}_2 = &\quad \mathbf{U}_2 \, \mathbf{\Sigma}_2 \, \mathbf{V}_2^{\mathrm{T}} . \end{aligned} \tag{3.25}$$

For Block i (= 1,2), the percentage variation captured by the k^{th} factor is:

$$p_{i,k}\% = 100 \frac{\mathbf{\Sigma}^2_{i\,(k,k)}}{\sum_{j=1}^{j=n_{d_i}} \mathbf{\Sigma}^2_{i\,(j,j)}} , \tag{3.26}$$

where, $\mathbf{\Sigma}_{i\,(j,j)}$ denotes the jth diagonal element of matrix $\mathbf{\Sigma}_i$, for $i = 1, 2$; $n_{d_1} = 62$, and $n_{d_2} = 248$. Let $\mathbf{v}_{1\,(k)}$ be the k^{th} column of $\mathbf{V}_1$. The projection of $\hat{\mathbf{D}}_2$ onto $\mathcal{C}(\mathbf{v}_{1\,(k)})$ is $\hat{\mathbf{D}}_2 \mathbf{v}_{1\,(k)} \mathbf{v}_{1\,(k)}^{\mathrm{T}}$, which can be factorized using SVD:

$$\hat{\mathbf{D}}_2 \mathbf{v}_{1\,(k)} \mathbf{v}_{1\,(k)}^{\mathrm{T}} = \mathbf{U}_3 \, \mathbf{\Sigma}_3 \, \mathbf{V}_3^{\mathrm{T}} . \tag{3.27}$$

Then, the percentage variation of $\hat{\mathbf{D}}_2$ in the direction of $\mathbf{v}_{1\,(k)}$ is:

$$p_{3,k}\% = 100\frac{\mathbf{\Sigma}^2_{3\,(1,1)}}{\sum_{j=1}^{j=248}\mathbf{\Sigma}^2_{2\,(j,j)}}\,. \tag{3.28}$$

The similarity of $p_{2,k}$ and $p_{3,k}$, for different values of k, indicates the similarity of the subspaces $\mathcal{R}(\hat{\mathbf{D}}_1)$ and $\mathcal{R}(\hat{\mathbf{D}}_2)$. This is shown in Table 3.1i. High percentage of variance captured using few components (for all cases more than 98% variation is captured with 3 components), and the close correspondence of $p_{2,k}$ and $p_{3,k}$, for $k = 1, \ldots, 5$, suggest that the drift space due to differences in instrumental responses is low dimensional. Similar results are obtained with Dataset 2 (see Table 3.1ii).

Using Dataset 3, $\hat{\mathbf{D}}$ due to moisture effects is computed using master and slave spectra at corresponding values of particle size and temperature, for each of the 11 samples, leading to $\hat{\mathbf{D}}$ of size $[264 \times 550]$. Choosing two samples randomly out of the eleven, $\hat{\mathbf{D}}$ is partitioned into two blocks $\hat{\mathbf{D}}_1$ and $\hat{\mathbf{D}}_2$, and the same procedure is repeated to obtain $p_{2,k}$ and $p_{3,k}$, for different values of k. The results are shown in Table 3.2i. $\hat{\mathbf{D}}$ due to particle size variation, and temperature variation are also found to be low dimensional (see Table 3.2ii and 3.2iii). With moisture, particle size, and temperature effects together, the results are shown in Table 3.2iv. In each case, the loadings computed from the first block of data correspond to the loadings computed for the second block of data, i.e. $p_{2,k} \approx p_{3,k}$, for $k = 1, \ldots, 5$. This suggests that drift due to a variety of different sources lies in a low-dimensional space.

(i)

k	$p_{2,k}\%$	$p_{3,k}\%$
1	91.36	91.22
2	5.17	5.15
3	2.00	2.03
4	0.77	0.77
5	0.34	0.37

(ii)

k	$p_{2,k}\%$	$p_{3,k}\%$
1	92.13	91.75
2	4.01	4.00
3	2.32	2.51
4	0.95	1.03
5	0.36	0.39

Table 3.1. First example: Study of drift due to differences in instrumental responses for (i) Dataset 1, and (ii) Dataset 2.

(i)

k	$p_{2,k}\%$	$p_{3,k}\%$
1	96.78	96.23
2	2.83	3.32
3	0.32	0.35
4	0.04	0.05
5	0.02	0.03

(ii)

k	$p_{2,k}\%$	$p_{3,k}\%$
1	99.46	99.43
2	0.30	0.28
3	0.20	0.25
4	0.01	0.01
5	0.01	0.01

(iii)

k	$p_{2,k}\%$	$p_{3,k}\%$
1	97.35	97.22
2	1.51	1.54
3	0.66	0.68
4	0.32	0.36
5	0.07	0.08

(iv)

k	$p_{2,k}\%$	$p_{3,k}\%$
1	97.14	96.96
2	2.62	2.77
3	0.18	0.20
4	0.03	0.03
5	0.02	0.02

Table 3.2. First example: Study of drift (in Dataset 3) due to (i) moisture, (ii) particle size variations, (iii) temperature, and (iv) moisture, particle size, and temperature variations together.

3.5.2 Second example (experimental data): Spectra measured at different temperatures

Calibration, prediction and master/slave data: NIR spectra of 22 mixtures of ethanol, water and isopropanol are measured at 5 temperatures (30, 40, 50, 60, 70 °C) on a HP 8453 UV-VIS spectrometer at 512 wavelengths in the range of 580–1091 nm [47]. The properties of interest are the mole fractions of the analytes. Fig. 3.2 shows the spectra of Sample#11 at five temperatures. Significant differences can be observed due to temperature variation.

For illustration purposes, the data are split in such a way that the prediction data contains drift unseen in the calibration data. This is done by choosing the calibration data and the prediction data at distinct temperatures and concentrations. Let $\mathbf{X}_a$ be the $[22 \times 512]$ spectra at temperature T_a, $a = 1, \ldots, 5$. For all a, let $\mathbf{X}_a$ be split into two $[11 \times 512]$ matrices, $\mathbf{X}_{a,o}$ and $\mathbf{X}_{a,e}$ consisting of odd and even numbered samples, respectively. Then, the $[22 \times 512]$ calibration spectra $\tilde{\mathbf{X}}_c$ and the $[55 \times 512]$ prediction spectra $\tilde{\mathbf{X}}_p$ are chosen to be:

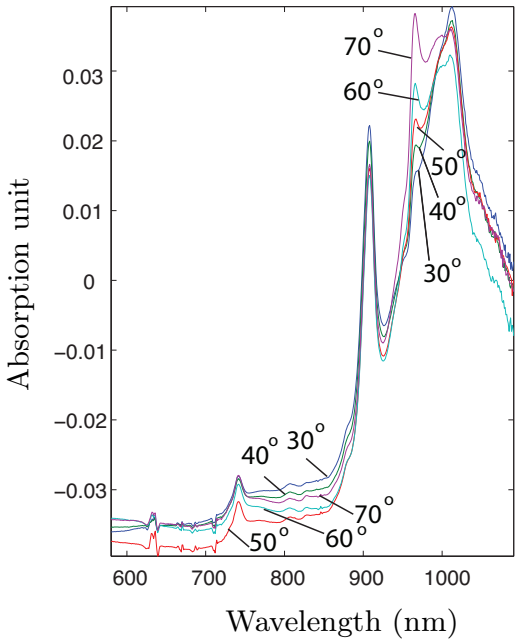


Fig. 3.2. Example 2: Spectra of Sample#11 at five temperatures.

$$\tilde{\mathbf{X}}_c = \begin{bmatrix} \mathbf{x}_{1,o} \\ \mathbf{x}_{2,o} \end{bmatrix}; \qquad \tilde{\mathbf{X}}_p = \begin{bmatrix} \mathbf{x}_{1,e} \\ \mathbf{x}_{2,e} \\ \mathbf{x}_{3,e} \\ \mathbf{x}_{4,e} \\ \mathbf{x}_{5,e} \end{bmatrix}. \tag{3.29}$$

Master/slave data of Types 1 and 2 are used for drift-space estimation. For both types, the slave matrix $\tilde{\mathbf{X}}_s$ is composed of two samples ($q = 2$) at five temperatures ($t = 5$) randomly chosen from the prediction spectra matrix $\tilde{\mathbf{X}}_p$ (i.e. $n_d = 10$). For Type 1 data, $\tilde{\mathbf{X}}_m$ is computed as in Eqs. (3.7)–(3.8), while for Type 2 data, $\tilde{\mathbf{X}}_m$ is computed using a minimum-norm solution to Eq. (3.10). Note that the concentrations of the Type 1 data can be unknown, while the concentrations of the three analytes of the Type 2 data must be known.

The drift subspace is estimated from Eq. (3.6). Drift is corrected using shrinkage and OP. Calibration models of column-mean centered data are then built using PLSR with the numbers of PLSR factors corresponding to the minimum RRMSEP^{app}. The RRMSEP^{app} is averaged over all combinations for selecting two samples out of the 11 samples at the 5 temperatures, i.e. $\binom{11}{2} = 55$ Monte Carlo simulations. The averaged RRMSEPs^{app} of the calibration models

without correction, with shrinkage, and with OP are compared for a range of values of the meta-parameters α and r_d.

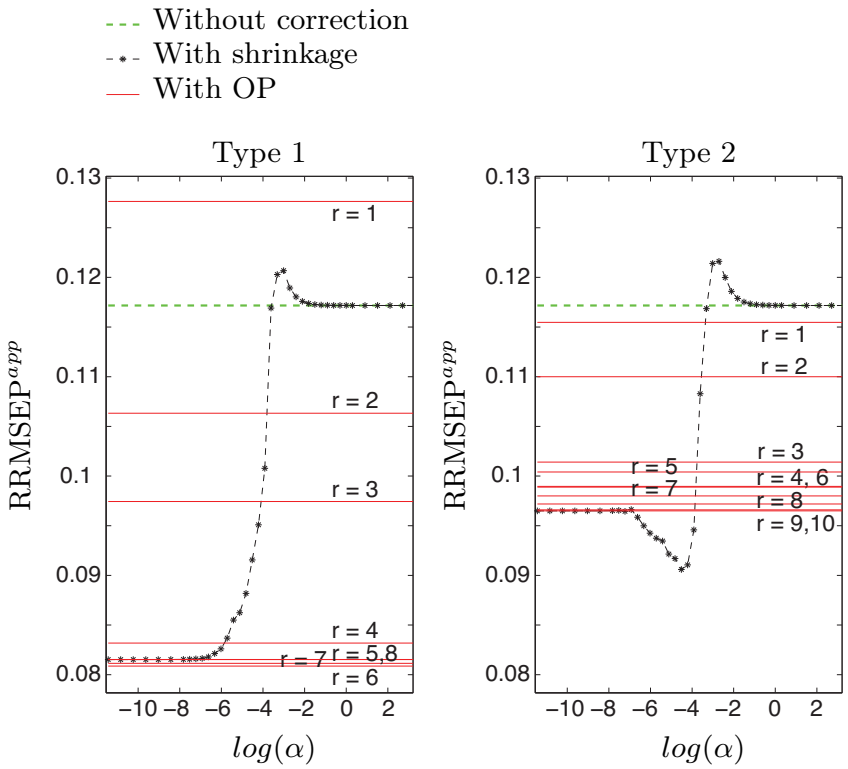


Fig. 3.3. Experimental Example 2: RRMSEPs^{app} averaged over 55 Monte Carlo simulations, without correction, with shrinkage, and with OP for a range of meta-parameter values and master/slave data of Types 1 and 2.

Discussion: Fig. 3.3 shows the RRMSEPs^{app} for ethanol without drift correction, with OP, and with shrinkage. The RRMSEP^{app} of shrinkage varies as a function of α, whereas separate horizontal lines for each value of $r_d = 1, \ldots, r_{max}$ denote the RRMSEPs^{app} with OP. For small α, RRMSEP^{app} of shrinkage is the same as RRMSEP^{app} of OP with $r_d = r_{max} = 8$ for Type 1 and $r_d = r_{max} = 10$ for Type 2. For large α, RRMSEP^{app} of shrinkage is the same as RRMSEP^{app} with no correction, thus confirming Proposition 1. It can be seen in Fig. 3.3 that Type 2 data lead to larger RRMSEPs^{app} than Type 1 data for most of the meta-parameter settings. The non-monotonicity in the RRMSEP^{app} of shrinkage with Type 2 data implies a significant contribution from error terms $e2$ and $e3$, possibly due to high overlap between the estimated drift and the signal.

While it is possible to get a better RRMSEP^{app} for certain values of α, realizing the gain is a matter of choosing the meta-parameters correctly, which may be difficult with small amounts of master/slave data.

3.5.3 Third example (experimental data): Spectra measured from samples collected at different plants

Calibration, prediction and master/slave data: UV spectra of 114 samples of light gas oil and diesel fuel are measured on a Cary 3 UV-VIS spectrometer at 572 wavelengths in the range of 200–400 nm [48]. The properties of interest are the weight percentages of saturates, monoaromatics, diaromatics, and polyaromatics.

The 114 samples are obtained from three pilot plants exhibiting different treatment, feed, catalyst, etc. and measured on the same instrument. For illustration purposes, the data are split in such a way that the prediction data contains drift unseen in the calibration data. This is done by choosing the calibration data and the prediction data from distinct pilot plants: the spectroscopic measurements from the 3rd pilot plant (30 samples) are used for calibration and those from the 1st and 2nd pilot plants (59 and 25 samples, respectively) are used for prediction. Let $\mathbf{X}_a$ be the spectra from plant a, $a = 1, 2, 3$. The $[30 \times 572]$ calibration spectra $\tilde{\mathbf{X}}_c$ and the $[84 \times 572]$ prediction spectra $\tilde{\mathbf{X}}_p$ are:

$$\tilde{\mathbf{X}}_c = \mathbf{X}_3; \qquad \tilde{\mathbf{X}}_p = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}. \tag{3.30}$$

Master/slave data of Type 2 are used for drift-space estimation. The slave matrix $\tilde{\mathbf{X}}_s$ is composed of three samples randomly chosen from $\mathbf{X}_1$ and two from $\mathbf{X}_2$ (i.e. $n_d = 5$). The master matrix $\tilde{\mathbf{X}}_m$ is computed using a minimum-norm solution to Eq. (3.10). Drift is corrected as in Example 2. The RRMSEP^{app} is averaged over 100 combinations for selecting the five correction samples out of the 84 prediction samples.

Discussion: Fig. 3.4 shows the RRMSEPs^{app} for percentage polyaromatics without drift correction, with OP, and with shrinkage. For small α, RRMSEP^{app} of shrinkage is the same as RRMSEP^{app} of OP with $r_d = r_{max} = 5$. For large α, RRMSEP^{app} of shrinkage is the same as RRMSEP^{app} with no correction. It can be seen in Fig. 3.4 that drift correction based on shrinkage is a monotonically

increasing function of α, implying that the drift term dominates the trade-off between the first three error terms in Eq. (3.14). Also, the correction with OP is not sensitive to the choice of r_d, implying that the drift has only one preferential direction.

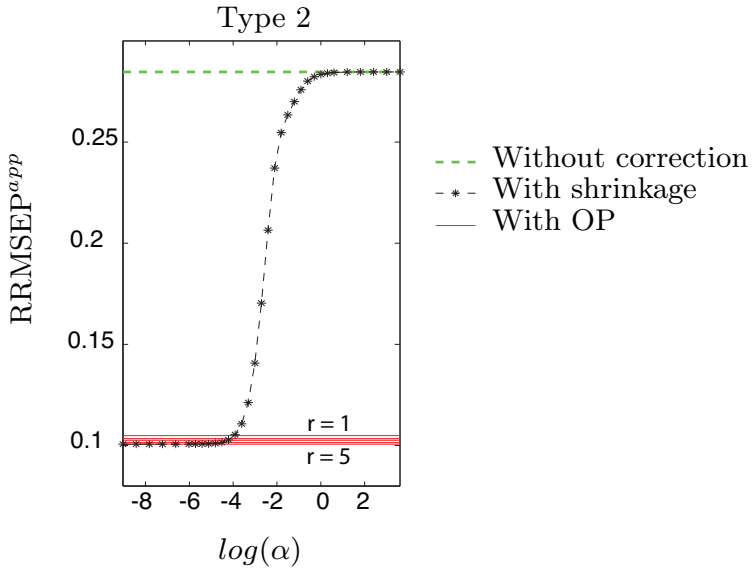


Fig. 3.4. Experimental Example 3: RRMSEPs^{app} averaged over 100 Monte Carlo simulations, without correction, with shrinkage, and with OP for a range of meta-parameter values and master/slave data of Type 2.

3.5.4 Fourth example (experimental data): Spectra measured with instrumental drift

Calibration, prediction and master/slave data: During anaerobic fermentation of glucose, Fourier-transform infrared (FTIR) spectra is monitored using a ReactIR$^{\text{TM}}$ 4000 single-beam spectrometer at 142 wavelengths in the frequency range of 1500–950 cm^{-1} [38]. The properties of interest are the molar concentrations of metabolites such as glucose, ethanol, ammonium, phosphates and glycerol. The FTIR spectra of $n_c = 49$ mixture samples constitute the $[49 \times 142]$ calibration spectra $\tilde{\mathbf{X}}_c$, and the spectra of $n_p = 16$ samples collected during a batch run constitute the $[16 \times 142]$ prediction spectra $\tilde{\mathbf{X}}_p$. High performance liquid chromatography (HPLC) concentration measurements of the 16 samples are used for validation.

Instrumental drift and physico-chemical drift is expected during batch operation due to changing spectral interactions between species. Master/slave data of Type 2 are used for drift-space estimation. Six-component mixtures are injected into the culture medium at roughly equal time intervals during the run (i.e. $n_d = 6$). The slave matrix $\tilde{\mathbf{X}}_s$ is composed of the difference in spectra before and after the injections. The master matrix $\tilde{\mathbf{X}}_m$ is computed using a minimum-norm solution to Eq. (3.10). Drift is corrected as in Example 2, except that the correction samples are fixed, i.e. the RRMSEP^{app} is not averaged over different combinations of correction samples.

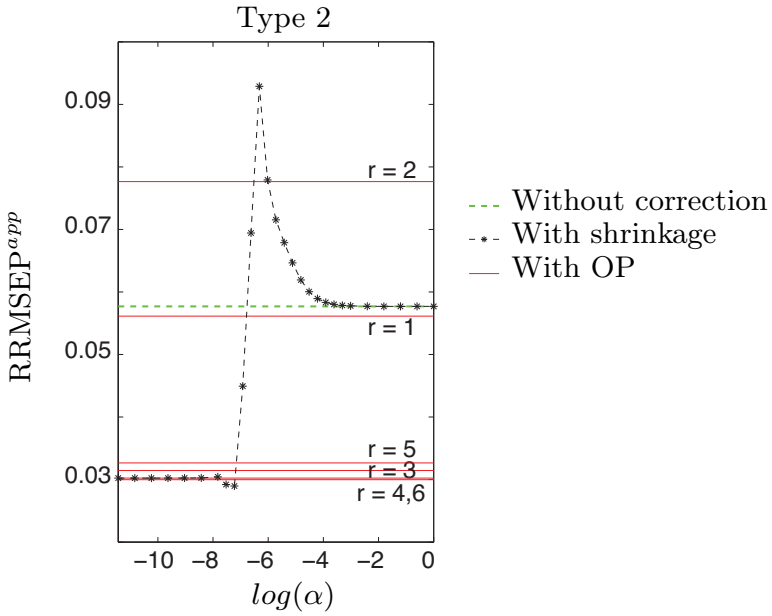


Fig. 3.5. Experimental Example 4: RRMSEPs^{app} without correction, with shrinkage, and with OP for a range of meta-parameter values and master/slave data of Type 2.

Discussion: Fig. 3.5 shows the RRMSEPs^{app} for molar concentration of glucose without drift correction, with OP, and with shrinkage. For small α, RRMSEP^{app} of shrinkage is the same as RRMSEP^{app} of OP with $r_d = r_{max} = 6$. For large α, RRMSEP^{app} of shrinkage is the same as RRMSEP^{app} with no correction. At the optimal values of meta-parameters, the RRMSEPs^{app} with shrinkage and with OP are within 0.2% RRMSEP^{app}. While it is possible to get a better RRMSEP^{app} with shrinkage and OP, a poor choice of the meta-parameters r_d and α can potentially lead to worse RRMSEPs^{app} than without correction (see

shrinkage with $10^{-6} < \alpha < 10^{-4}$ and OP with $r_d = 2$). The sharp peak observed in the RRMSEP^{app} of shrinkage cannot be validated as an artifact due to noise since the number of prediction samples n_p is too small.

3.5.5 Fifth example (experimental data): Spectra measured using different instruments

Calibration, prediction and master/slave data: NIR spectra of 80 samples of corn are measured on 3 spectrometers at 700 wavelengths in the range of 1100–2498 nm [49]. The properties of interest are the weight percentages of moisture, oil, protein and starch.

For illustration purposes, the data are split in such a way that the prediction data contains drift unseen in the calibration data. This is done by choosing the calibration data and the prediction data from distinct instruments and at distinct concentrations: Let $\mathbf{X}_a$ be the $[80 \times 700]$ spectra on instrument I_a, $a = 1, 2$. For both a, let $\mathbf{X}_a$ be split into two $[40 \times 700]$ matrices, $\mathbf{X}_{a,o}$ and $\mathbf{X}_{a,e}$ consisting of odd and even numbered samples, respectively. Then, the $[40 \times 700]$ calibration spectra $\tilde{\mathbf{X}}_c$ and the $[40 \times 700]$ prediction spectra $\tilde{\mathbf{X}}_p$ are chosen to be:

$$\tilde{\mathbf{X}}_c = \mathbf{X}_{1,o}; \qquad \tilde{\mathbf{X}}_p = \mathbf{X}_{2,e}. \tag{3.31}$$

For calibration transfer, typically Type 1 data are used. However, in this example it is shown that Type 2 can also be used for calibration transfer. For Type 1, five samples ($q = 5$) randomly chosen from the forty even prediction samples are measured on the two instruments ($t = 2$), leading to $n_d = 10$. It can easily be shown for this two-instrument case that $\tilde{\mathbf{X}}_s = \begin{bmatrix} \mathbf{X}_{s,1} \\ \mathbf{X}_{s,2} \end{bmatrix}$, $\tilde{\mathbf{X}}_m = \begin{bmatrix} (\mathbf{X}_{s,1}+\mathbf{X}_{s,2})/2 \\ (\mathbf{X}_{s,1}+\mathbf{X}_{s,2})/2 \end{bmatrix}$ can be simplified to $\tilde{\mathbf{X}}_s = \mathbf{X}_{s,2}$, $\tilde{\mathbf{X}}_m = \mathbf{X}_{s,1}$. For Type 2, the slave matrix $\tilde{\mathbf{X}}_s$ is composed of the five spectroscopic measurements on the second instrument, while $\tilde{\mathbf{X}}_m$ is computed using a minimum-norm solution to Eq. (3.10). Hence, the procedure for Type 2 requires the samples to be measured on only one instrument (i.e. $n_d = 5$). Drift is corrected as in Example 2. The RRMSEP^{app} is averaged over 100 combinations for selecting five correction samples out of the 40 prediction samples.

Discussion: Fig. 3.6 shows the RRMSEPs^{app} for moisture content without drift correction, with OP, and with shrinkage. For small α, RRMSEP^{app} of

shrinkage is the same as RRMSEP^{app} of OP with $r_d = r_{max} = 5$ for Type 1 and Type 2. For large α, RRMSEP^{app} of shrinkage is the same as RRMSEP^{app} with no correction. As in the previous examples, it can be inferred that the non-monotonicity in the RRMSEP^{app} of shrinkage implies a significant contribution from error terms $e2$ and $e3$, possibly due to high overlap between the estimated drift and the signal. In contrast to Example 2, in this example Type 2 data lead to smaller RRMSEPs^{app} than Type 1 data. Note also that it is possible to get a better RRMSEP^{app} in some cases with shrinkage ($\alpha \simeq 10^{-3}$ for Type 1, $\alpha \simeq 10^{-4}$ for Type 2) than with OP with any r_d. However, at the optimal values of meta-parameters, the RRMSEPs^{app} with shrinkage and with OP are within 0.5% RRMSEP^{app}. Furthermore, a good value of α is not known *a priori* and is difficult to determine with small amounts of master/slave data.

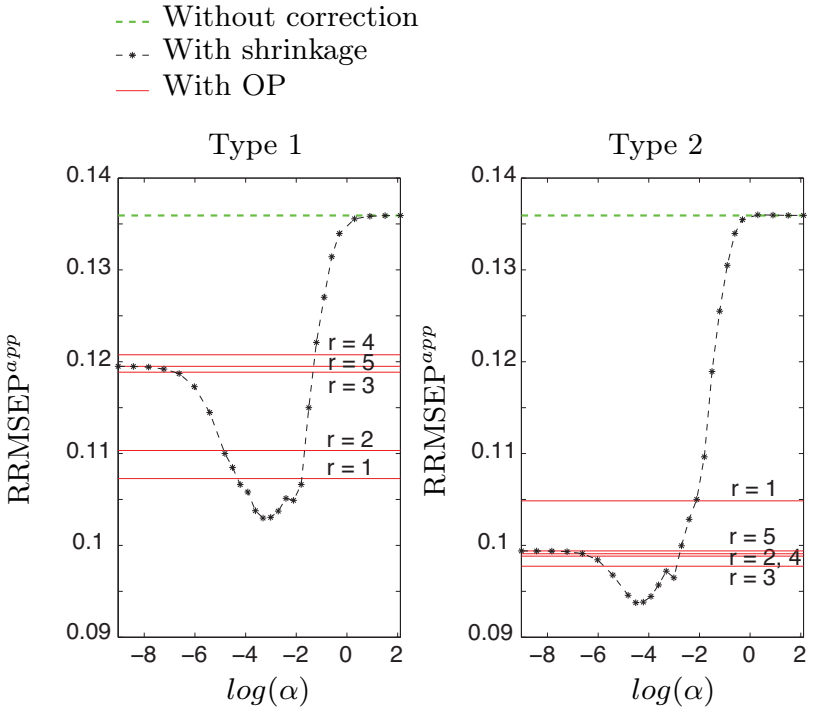


Fig. 3.6. Experimental Example 5: RRMSEPs^{app} averaged over 100 Monte Carlo simulations, without correction, with shrinkage, and with OP for a range of meta-parameter values and for master/slave data of Types 1 and 2.

3.6 Related methods

The scope of the chapter did not include related methods discussed in the following four points.

1. A measurement-based drift-correction method may pre-process the spectra (*predictor correction*) or post-process the concentrations (*predictand correction*). The chapter has investigated only predictor-correction methods. Examples of predictand-correction methods include Kalman filtering for slope-and-bias correction [50, 51] using on-line master/slave data, and process migration [52] using an additional predictand-correction model built off-line. However, Kalman filtering for slope-and-bias correction requires frequent master/slave data to track the drift locally, while process migration requires many representative correction samples to build a global off-line model.

2. For forward calibration, Haaland and Melgaard proposed a prediction-augmented classical least squares (PACLS) [53]. However, the PACLS algorithm is often too restrictive since the shapes of all the drift components need to be known [54]. To overcome this restriction, the same authors proposed a hybrid PACLS/PLS algorithm that proceeds in two steps. In Step 1, the known shapes of the drift components are modeled using PACLS. In Step 2, the unknown shapes of the drift components are extracted from the residues of the PACLS model using PLSR. If the goal of fusing forward and inverse calibration modeling was better interpretability, the hybrid PACLS/PLSR defeats its purpose since the pure components estimated in Step 1 are corrupted by the unknown drift, thereby leading to poor interpretability. Hence, despite the tedious modeling of known drift, the hybrid PACLS/PLS seems to offer little advantage over drift-correction methods for inverse calibration.

3. A calibration model based on orthogonal partial least squares (OPLS) has an inbuilt filter for orthogonal signal correction (OSC), that can be used to estimate the drift subspace [55, 56]. However, OPLS suffers from the drawbacks of ICM. Assuming that the amount of slave data (with drift) is much smaller compared to the original calibration data (with no drift), building an OPLS calibration model on the augmented data will not pro-

vide good estimates of the drift subspace, since small amounts of slave data will typically account for a small percentage of the variation in the augmented data. While OPLS often results in a model with fewer factors, its prediction accuracy is similar to that obtained with standard PLSR built on the augmented data [57].

4. Another solution for drift correction is to use prior knowledge to build a drift-invariant calibration model. For example, polynomial principal component regression (polyPCR) and pseudo principal component regression (pPCR) models are invariant to all polynomial-shape drifts of a user-specified order [58]; data pretreatment via standard normal variate (SNV) or multiplicative scatter correction (MSC) corrects specifically for possible scatter effects [59], and first-order or second-order differentiation renders prediction invariant to offsets that are constant or linear along the spectrum. However, the drift model must be known beforehand, which is seldom the case.

3.7 Conclusions

This chapter provides a framework for ECMs that consists of two main steps: (i) estimation of the drift subspace based upon different types of master/slave data, and (ii) correction of the calibration model for the estimated drift subspace by shrinkage or orthogonal projection. The drift subspace is estimated in a master/slave setting, whereby the master/slave data are measured for the slave with drift and computed for the master (with no drift) through a linear operator. This work characterizes the master/slave data as Type 1 if $\mathbf{A}$ is deduced from the knowledge of the master spectra (e.g., mean-centered spectra), and as Type 2 if $\mathbf{A}$ is computed from the known master concentrations (e.g., from reference measurements or prior knowledge) and the calibration concentrations. It has been shown analytically and with experimental examples that, for large α, shrinkage corresponds to no correction, while drift correction by OP can be seen as a special case of shrinkage when $r_d = \text{rank}\left(\hat{\mathbf{D}}\right)$ and small α, i.e. the drift subspace is shrunk completely. Lastly, in the absence of noise, the concept of pNAS is used to show that ICM and ECM with OP are equivalent.

4

Latent subspace correction using unlabeled data

This chapter studies latent subspace correction (to reduce RRMSEP_{e1}) based on additional information in the form of unlabeled data.

4.1 Overview

In LV calibration based on PCR or PLSR, most often, calibration data consist of *labeled data* only, i.e. X-measurements for which the corresponding y-measurements are available. In many practical applications, the set of labeled data is small owing to the high costs of measuring y-values. However, a large number of X-measurements may be available (so-called *unlabeled data*), which may be used together with the labeled data for calibration [60]. The use of labeled and unlabeled data for regression and classification modeling, commonly referred to as *semi-supervised learning*, represents an increasing focus of the applied statistics, machine learning, and chemometric literature. Several methods that improve prediction performance by a judicious use of labeled and unlabeled data, e.g. transductive support vector machine (SVM) [61], co-training [62], graph-based methods [63], imputation and expectation-maximization algorithm [64], have been developed. The present work focusses on the workhorses of chemometrics, namely PCR and PLSR.

Isaksson *et al.* and Thomas have developed a method, abbreviated here as IT-PCR, that uses both labeled and unlabeled data in the PCA step to stabilize the latent subspace in PCR [65, 66]. Ergon and Esbensen approached the

unlabeled data problem using optimal filtering (OF) [67, 68] and developed an OF-based PCR predictor that was shown to be equivalent to IT-PCR. The same authors also derived an OF-based predictor for PLSR. The loadings of this predictor are not estimated *sequentially*[1]. Besides being computationally intensive during cross-validation, non-sequential estimation leads to confounding between the latent vectors, and are hence difficult to interpret. This work proposes a sequential version of this OF-based PLSR that solves these problems through the addition of a deflation step. It is shown analytically that the sequential version of the OF-based PLSR is equivalent to a PLSR model built on the PCA scores estimated from the labeled and unlabeled data (PCA-PLSR).

Simulated and experimental data sets are used to point out the usefulness and pitfalls of using unlabeled data. By stabilizing the latent subspaces in the calibration step, these methods lead to a lower RMSEP. In other words, unlabeled data help reduce the amount of labeled data required for a particular RMSEP. Since labeled data are often expensive, while unlabeled data are sometimes freely available, there may be a cost incentive in using unlabeled data. It is shown experimentally that the advantage gained from using unlabeled data can be significant when the measurements have low SNR. This may be relevant, for example, for calibrating APIs present in low concentrations, or in NIR reflectance spectrum at high wavelengths due to light losses resulting from transport along the fiber-optic probes [3].

In regression modeling, it is generally assumed that the labeled data are representative of the prediction data. This may not be true in the presence of drift (see Chapter 3). It is shown with Monte Carlo simulations that, in the presence of drift, the use of unlabeled data can result in an increase in prediction error compared to that obtained with a model based on labeled data alone. The four constituents of the prediction error (see Eq. (2.26)) are analyzed separately, leading to a better understanding of the different effects in the presence of drift.

The chapter is organized as follows. The PCA-based and the OF-based use of unlabeled data, and the proposed extensions, are presented in Section 4.2. Section 4.3 discusses two propositions on the equivalence of various methods. Section 4.4 compares these methods for simulated and real spectroscopic data, and Section 4.5 concludes the chapter.

[1] Sequential estimation involves determining one loading followed by deflation of the measurement matrix, from which the next loading can be estimated.

4.2 Reducing subspace modeling error using unlabeled data

Let n_e measurements of unlabeled data $\tilde{\mathbf{X}}_e$ be available to help reduce the subspace modeling error. It is assumed that $\mathbf{X}_e$ is representative of $\mathbf{X}_p$ with respect to row-space, drift, and range of response variable.

4.2.1 PCA-based use of unlabeled data

Since the y-values are not required in PCA, it is reasonable to stabilize the principal component subspace using both the labeled and unlabeled data ($\tilde{\mathbf{X}}_c$ and $\tilde{\mathbf{X}}_e$) as discussed next.

PCR with unlabeled data (IT-PCR)

Isaksson *et al.* [65] and Thomas [66] proposed a PCR model with unlabeled data that uses the following steps:

1. Compute the PCA factorization of $\begin{bmatrix}\tilde{\mathbf{X}}_c \\ \tilde{\mathbf{X}}_e\end{bmatrix}$ and retain r_{PCR} factors:

$$\begin{bmatrix}\tilde{\mathbf{X}}_c \\ \tilde{\mathbf{X}}_e\end{bmatrix} = \mathbf{T}_{\text{IT}-\text{PCR}}\mathbf{P}_{\text{IT}-\text{PCR}}^{\text{T}} + \mathbf{E}_{\text{IT}-\text{PCR}}\,, \tag{4.1}$$

   where $\mathbf{T}_{\text{IT}-\text{PCR}}$, $\mathbf{P}_{\text{IT}-\text{PCR}}$ and $\mathbf{E}_{\text{IT}-\text{PCR}}$ are the $[(n_c + n_e) \times r_{\text{PCR}}]$ scores matrix, $[n_x \times r_{\text{PCR}}]$ loading matrix, and $[(n_c + n_e) \times n_x]$ residual matrix, respectively. $\mathbf{T}_{\text{IT}-\text{PCR}}$ can be decomposed as

$$\mathbf{T}_{\text{IT}-\text{PCR}} = \begin{bmatrix}\mathbf{T}_{\text{IT}-\text{PCR}-\text{l}} \\ \mathbf{T}_{\text{IT}-\text{PCR}-\text{u}}\end{bmatrix}\,, \tag{4.2}$$

   where $\mathbf{T}_{\text{IT}-\text{PCR}-\text{l}}$ and $\mathbf{T}_{\text{IT}-\text{PCR}-\text{u}}$ are the $[n_c \times r_{\text{PCR}}]$ and $[n_e \times r_{\text{PCR}}]$ scores matrices corresponding to the labeled and unlabeled data, respectively.

2. Compute $\hat{\mathbf{b}}$ from the least-squares regression between $\{\mathbf{T}_{\text{IT}-\text{PCR}-\text{l}}, \tilde{\mathbf{y}}_c\}$:

$$\hat{\mathbf{b}} = \mathbf{P}_{\text{IT}-\text{PCR}}(\mathbf{T}_{\text{IT}-\text{PCR}-\text{l}}^{\text{T}}\mathbf{T}_{\text{IT}-\text{PCR}-\text{l}})^{-1}\mathbf{T}_{\text{IT}-\text{PCR}-\text{l}}^{\text{T}}\tilde{\mathbf{y}}_c\,. \tag{4.3}$$

PCA-based PLSR (PCA-PLSR)

A direct extension of the above method for PLSR is proposed here with PCA-PLSR. PCA-PLSR uses both labeled and unlabeled data to stabilize the principal components as in Eq. (4.1), and then builds a PLSR model with r_{PLSR} factors on the data pair $\{\mathbf{T}_{\text{IT}-\text{PCR}-\text{l}}, \tilde{\mathbf{y}}_c\}$. The advantage of this approach is that typically fewer factors are required in PCA-PLSR than in IT-PCR, i.e. $r_{\text{PLSR}} \leq r_{\text{PCR}}$. A schematic block-diagram of PCA-PLSR is shown in Fig. 4.1.

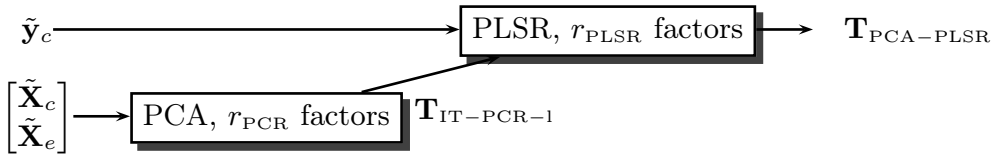


Fig. 4.1. Schematic block-diagram of PCA-PLSR

4.2.2 OF-based use of unlabeled data

Consider the LV model in Eq. (2.1). An optimal estimate of the scores $\hat{\mathbf{t}}$ is obtained by minimizing $\mathcal{E}[(\mathbf{t}-\hat{\mathbf{t}})(\mathbf{t}-\hat{\mathbf{t}})^{\text{T}}]$ over an $[r \times n_x]$ matrix $\mathbf{K}$, where $\hat{\mathbf{t}} = \mathbf{K}\,\tilde{\mathbf{x}}$. Differentiating $\mathcal{E}[(\mathbf{t}-\hat{\mathbf{t}})(\mathbf{t}-\hat{\mathbf{t}})^{\text{T}}]$ with respect to $\mathbf{K}$ and equating to zero leads to:

$$\mathbf{K} = \mathbf{R}_t\mathbf{L}^{\text{T}}(\mathbf{L}\mathbf{R}_t\mathbf{L}^{\text{T}} + \mathbf{R}_v)^{-1}\,, \tag{4.4}$$

where $\mathbf{R}_t = \mathcal{E}[\mathbf{t}\,\mathbf{t}^{\text{T}}]$, and $\mathbf{R}_v = \mathcal{E}[\mathbf{v}_x\,\mathbf{v}_x^{\text{T}}]$ (see [67] for details). $\mathbf{K}$ can be interpreted as a regularized inverse of $\mathbf{L}$.

The OF-based PCR predictor has been shown to be equivalent to IT-PCR [67] and is discussed in Appendix B.1. Only the OF-based PLSR is presented in the following sections.

Non-sequential optimized PLSR (NSO-PLSR)

The main steps proposed in [67, 68] are repeated below:

1. Compute the scores and loadings based on Martens' PLSR [19, 20] using the data pair $\{\tilde{\mathbf{X}}_c, \tilde{\mathbf{y}}_c\}$ and retaining r_{PLSR} factors,

$$\begin{aligned} \tilde{\mathbf{X}}_c &= \mathbf{T}_{\text{M}}\mathbf{W}^{\text{T}} + \mathbf{E}_{\text{M}} \\ \tilde{\mathbf{y}}_c &= \mathbf{T}_{\text{M}}\mathbf{q}_{\text{M}} + \mathbf{f}\,. \end{aligned} \tag{4.5}$$

   Martens' PLSR is chosen over the more popular Wold's version of PLSR because Martens' PLSR leads to estimates of $\mathbf{T}_{\text{M}}$, $\mathbf{W}$ and $\mathbf{E}_{\text{M}}$ that are consistent with the OF-based PLSR (see Appendix B.2).

2. Compute the OF-based scores $\mathbf{T}_{\text{NSO}-\text{PLSR}} = \tilde{\mathbf{X}}_c\hat{\mathbf{K}}^{\text{T}}$, where

$$\hat{\mathbf{K}} = \hat{\mathbf{R}}_t\mathbf{W}^{\text{T}}(\mathbf{W}\hat{\mathbf{R}}_t\mathbf{W}^{\text{T}} + \hat{\mathbf{R}}_v)^{-1}\,, \tag{4.6}$$

   and $\hat{\mathbf{R}}_t$ and $\hat{\mathbf{R}}_v$ are the empirical estimates of the scores and X-noise covariance matrices, respectively. $\hat{\mathbf{R}}_t$ is computed using PLSR scores:

$$\hat{\mathbf{R}}_t = \frac{1}{n_c - 1}\mathbf{T}_{\text{M}}^{\text{T}}\mathbf{T}_{\text{M}}\,. \tag{4.7}$$

   Similarly, $\hat{\mathbf{R}}_v$ can be computed using PLSR X-residuals, i.e. $\hat{\mathbf{R}}_v = \frac{1}{n_c-1}\mathbf{E}_{\text{M}}^{\text{T}}\mathbf{E}_{\text{M}}$. However, the PLSR X-residuals lead to a poor estimate of the true X-noise covariance matrix [21]. Moreover, the PLSR models built separately for each analyte lead to different estimates of $\hat{\mathbf{R}}_v$. On the other hand, PCA has been shown useful to compute the X-noise covariance matrix [21]. Furthermore, an additional advantage of PCA is that both labeled and unlabeled data can be used. This leads to:

$$\hat{\mathbf{R}}_v = \frac{1}{n_c + n_e - 1}\mathbf{E}_{\text{IT}-\text{PCR}}^{\text{T}}\mathbf{E}_{\text{IT}-\text{PCR}}\,, \tag{4.8}$$

   where $\mathbf{E}_{\text{IT}-\text{PCR}}$ is defined in Eq. (4.1).

3. Compute $\hat{\mathbf{b}}$ from the least-squares regression between $\{\mathbf{T}_{\text{NSO}-\text{PLSR}}, \tilde{\mathbf{y}}_c\}$:

$$\hat{\mathbf{b}} = \hat{\mathbf{K}}^{\text{T}}(\mathbf{T}_{\text{NSO}-\text{PLSR}}^{\text{T}}\mathbf{T}_{\text{NSO}-\text{PLSR}})^{-1}\mathbf{T}_{\text{NSO}-\text{PLSR}}^{\text{T}}\tilde{\mathbf{y}}_c\,. \tag{4.9}$$

This method will be referred to as non-sequential optimized PLSR (NSO-PLSR) since the loadings are estimated non-sequentially. This leads to confounding in

latent vectors, meaning that the scores (or loading) vector with $r_{\text{PLSR}} = 1$ is not the same as the first of the two scores (or loading) vectors with $r_{\text{PLSR}} = 2$. A schematic block-diagram of NSO-PLSR is shown in Fig. 4.2.

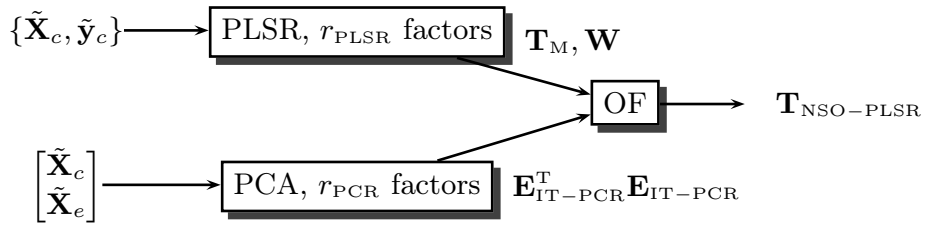


Fig. 4.2. Schematic block-diagram of NSO-PLSR

Sequential optimized PLSR (SO-PLSR)

Non-sequential fitting methods involve heavy computations during cross-validation since higher factors cannot be computed incrementally. Furthermore, NSO-PLSR is difficult to interpret for two reasons: (i) neither the loading matrix $\hat{\mathbf{K}}$ in Eq. (4.6) nor the scores $\mathbf{T}_{\text{NSO-PLSR}}$ are orthogonal, and (ii) the scores and loadings are confounded. We propose next a sequentially-estimated OF-based PLSR that solves these problems. This method is a minor adaptation of NSO-PLSR, as it proceeds with the same Steps 1–3, except that one scores vector is computed at a time, followed by deflation to ensure orthogonality with the subsequent scores vectors.

The schematic block-diagram of SO-PLSR is shown in Fig. 4.3, where $\{\mathbf{X}_{c*}(1), \mathbf{y}_{c*}(1)\} = \{\tilde{\mathbf{X}}_c, \tilde{\mathbf{y}}_c\}$, and deflation is defined as:

$$\mathbf{X}_{c*}(k+1) = \left(\mathbf{I} - \frac{\mathbf{t}_{\text{SO-PLSR}}(k)\mathbf{t}_{\text{SO-PLSR}}(k)^{\text{T}}}{\mathbf{t}_{\text{SO-PLSR}}(k)^{\text{T}}\mathbf{t}_{\text{SO-PLSR}}(k)}\right)\mathbf{X}_{c*}(k)\,. \tag{4.10}$$

The steps are repeated until r_{PLSR} scores are computed to obtain $\mathbf{T}_{\text{SO-PLSR}}$ of size $[n_c \times r_{\text{PLSR}}]$. Note that, at each iteration, $\hat{R}_t = \frac{1}{n_c-1}\mathbf{t}_{\text{M}}^{\text{T}}\mathbf{t}_{\text{M}}$ is a scalar, and $\hat{\mathbf{k}}^{\text{T}} = \hat{R}_t\,\mathbf{w}^{\text{T}}(\mathbf{w}\,\hat{R}_t\,\mathbf{w}^{\text{T}} + \hat{\mathbf{R}}_v)^{+}$ is a vector.

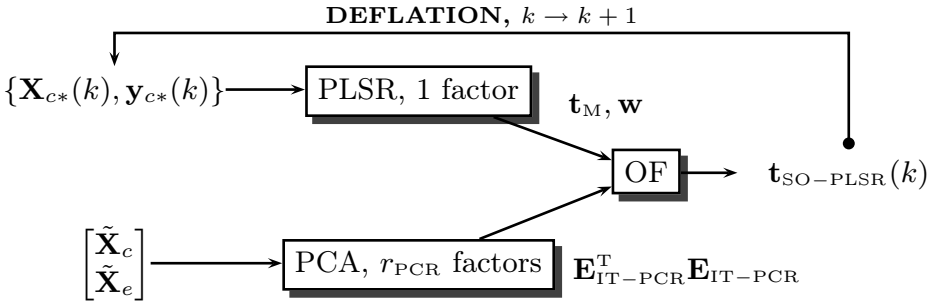


Fig. 4.3. Schematic block-diagram of SO-PLSR

4.3 Equivalence of methods

This section presents two theoretical results about the equivalence of different regression methods with unlabeled data.

4.3.1 Proposition 3: Equivalence of SO-PLSR and PCA-PLSR

Proposition 3 *SO-PLSR and PCA-PLSR lead to the same scores and loadings.*

Proof:
First, we present two lemmas that will be used in the proof.

Lemma 1 *Let* $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{B} \in \mathbb{R}^{n\times l}$, $\mathbf{C} \in \mathbb{R}^{k\times m}$, *and* $\mathbf{D} \in \mathbb{R}^{k\times l}$. *Then,*

$$\begin{bmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}\end{bmatrix} = \begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{C}\mathbf{A}^{+} & \mathbf{I}\end{bmatrix} \begin{bmatrix}\mathbf{A} & \mathbf{B}-\mathbf{A}\mathbf{A}^{+}\mathbf{B} \\ \mathbf{C}-\mathbf{C}\,\mathbf{A}^{+}\mathbf{A} & \mathbf{D}-\mathbf{C}\mathbf{A}^{+}\mathbf{B}\end{bmatrix} \begin{bmatrix}\mathbf{I} & \mathbf{A}^{+}\mathbf{B} \\ \mathbf{0} & \mathbf{I}\end{bmatrix} \tag{4.11}$$

Proof of Lemma 1: (see page 237 in [69])

Lemma 2 *Let* $\mathbf{a}_1 \in \mathbb{R}^n$ *(*$\mathbf{a}_1 \neq \mathbf{0}_n$*),* $\mathbf{a}_2 \in \mathbb{R}^m$ *(*$\mathbf{a}_2 \neq \mathbf{0}_m$*), and* $\mathbf{C} \in \mathbb{R}^{m\times m}$ *a diagonal matrix with* $\operatorname{rank}(\mathbf{C}) = m$. *Then,*

$$\begin{bmatrix}\mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_1\mathbf{a}_2^{\mathrm{T}} \\ \mathbf{a}_2\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_2\mathbf{a}_2^{\mathrm{T}}+\mathbf{C}\end{bmatrix}^{+} = \begin{bmatrix}\frac{(1+\mathbf{a}_2^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{a}_2)\mathbf{a}_1\mathbf{a}_1^{\mathrm{T}}}{||\mathbf{a}_1||^4} & \frac{-\mathbf{a}_1\mathbf{a}_2^{\mathrm{T}}\mathbf{C}^{-1}}{||\mathbf{a}_1||^2} \\ \frac{-\mathbf{C}^{-1}\mathbf{a}_2\mathbf{a}_1^{\mathrm{T}}}{||\mathbf{a}_1||^2} & \mathbf{C}^{-1}\end{bmatrix} \tag{4.12}$$

Proof of Lemma 2:
Using Lemma 1, it can be shown that

$$\begin{bmatrix} \mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_1\mathbf{a}_2^{\mathrm{T}} \\ \mathbf{a}_2\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_2\mathbf{a}_2^{\mathrm{T}} + \mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \frac{\mathbf{a}_2\mathbf{a}_1^{\mathrm{T}}}{||\mathbf{a}_1||^2} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \frac{\mathbf{a}_1\mathbf{a}_2^{\mathrm{T}}}{||\mathbf{a}_1||^2} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} . \tag{4.13}$$

Hence,

$$\begin{bmatrix} \mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_1\mathbf{a}_2^{\mathrm{T}} \\ \mathbf{a}_2\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_2\mathbf{a}_2^{\mathrm{T}} + \mathbf{C} \end{bmatrix}^{+} = \begin{bmatrix} \mathbf{I} & \frac{\mathbf{a}_1\mathbf{a}_2^{\mathrm{T}}}{||\mathbf{a}_1||^2} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}^{+} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \frac{\mathbf{a}_2\mathbf{a}_1^{\mathrm{T}}}{||\mathbf{a}_1||^2} & \mathbf{I} \end{bmatrix}^{-1} . \tag{4.14}$$

It can be verified by direct substitution that $\begin{bmatrix} \mathbf{I} & \frac{\mathbf{a}_1\mathbf{a}_2^{\mathrm{T}}}{||\mathbf{a}_1||^2} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & -\frac{\mathbf{a}_1\mathbf{a}_2^{\mathrm{T}}}{||\mathbf{a}_1||^2} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$, $\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \frac{\mathbf{a}_2\mathbf{a}_1^{\mathrm{T}}}{||\mathbf{a}_1||^2} & \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\frac{\mathbf{a}_2\mathbf{a}_1^{\mathrm{T}}}{||\mathbf{a}_1||^2} & \mathbf{I} \end{bmatrix}$, and $\begin{bmatrix} \mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}^{+} = \begin{bmatrix} \frac{\mathbf{a}_1\mathbf{a}_1^{\mathrm{T}}}{(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1)^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{-1} \end{bmatrix}$. Substituting these into Eq. (4.14) leads to Eq. (4.12). □

Let the PCA factorization of $\begin{bmatrix} \tilde{\mathbf{X}}_c \\ \tilde{\mathbf{X}}_e \end{bmatrix}$ be

$$\begin{bmatrix} \tilde{\mathbf{X}}_c \\ \tilde{\mathbf{X}}_e \end{bmatrix} = \mathbf{T}_1\mathbf{P}_1^{\mathrm{T}} + \mathbf{T}_2\mathbf{P}_2^{\mathrm{T}} , \tag{4.15}$$

where $\mathbf{T}_1$ and $\mathbf{T}_2$ are the $[(n_c+n_e)\times r_{\mathrm{PCR}}]$ and $[(n_c+n_e)\times min(n_c+n_e, n_x) - r_{\mathrm{PCR}}]$ scores matrices, respectively, $\mathbf{P}_1$ and $\mathbf{P}_2$ the $[n_x \times r_{\mathrm{PCR}}]$ and $[n_x \times min(n_c+n_e, n_x) - r_{\mathrm{PCR}}]$ loading matrices, respectively. Consider the first loading vector $\mathbf{w}$ in PLSR on the data pair $\{\tilde{\mathbf{X}}_c, \tilde{\mathbf{y}}_c\}$. Since $\mathbf{w} \in \mathcal{R}\left(\begin{bmatrix} \mathbf{P}_1^{\mathrm{T}} \\ \mathbf{P}_2^{\mathrm{T}} \end{bmatrix}\right)$, let

$$\mathbf{w} := \mathbf{P}_1\mathbf{a}_1 + \mathbf{P}_2\mathbf{a}_2 . \tag{4.16}$$

Using $\mathbf{w} \propto \tilde{\mathbf{X}}_c^{\mathrm{T}}\tilde{\mathbf{y}}_c$ (see NIPALS in Section 2.2.3), the first scores vector in PLSR, $\mathbf{t}_{\mathrm{PLSR}} = \tilde{\mathbf{X}}_c\,\mathbf{w} \propto \tilde{\mathbf{X}}_c\tilde{\mathbf{X}}_c^{\mathrm{T}}\tilde{\mathbf{y}}_c$. Similarly, the first scores vector in PCA-PLSR, $\mathbf{t}_{\mathrm{PCA-PLSR}}$

$$\begin{aligned}\mathbf{t}_{\mathrm{PCA-PLSR}} &\propto (\tilde{\mathbf{X}}_c\mathbf{P}_1)\,(\tilde{\mathbf{X}}_c\mathbf{P}_1)^{\mathrm{T}}\tilde{\mathbf{y}}_c\,,\\ &\propto \tilde{\mathbf{X}}_c\mathbf{P}_1\mathbf{P}_1^{\mathrm{T}}\mathbf{w}\,. \end{aligned} \tag{4.17}$$

The definition of $\mathbf{w}$ in Eq. (4.16) leads to

$$\mathbf{t}_{\mathrm{PCA-PLSR}} \propto \tilde{\mathbf{X}}_c\mathbf{P}_1\mathbf{P}_1^{\mathrm{T}}\mathbf{w} = \tilde{\mathbf{X}}_c\mathbf{P}_1\mathbf{a}_1\,. \tag{4.18}$$

The first scores vector in SO-PLSR, $\mathbf{t}_{\mathrm{SO-PLSR}}$

$$\begin{aligned}\mathbf{t}_{\mathrm{SO-PLSR}} &= \tilde{\mathbf{X}}_c\hat{\mathbf{k}} = \tilde{\mathbf{X}}_c(\mathbf{w}\,\hat{R}_t\,\mathbf{w}^{\mathrm{T}} + \hat{\mathbf{R}}_v)^{+}\hat{R}_t\,\mathbf{w}\,,\\ &= \tilde{\mathbf{X}}_c(\mathbf{w}\mathbf{w}^{\mathrm{T}} + \hat{\mathbf{R}}_v/\hat{R}_t)^{+}\mathbf{w}\,. \end{aligned} \tag{4.19}$$

Let $\mathbf{w}_{\mathrm{SO-PLSR}} := (\mathbf{w}\mathbf{w}^{\mathrm{T}} + \hat{\mathbf{R}}_v/\hat{R}_t)^{+}\mathbf{w}$, and $\mathbf{C} := \mathbf{T}_2^{\mathrm{T}}\mathbf{T}_2/\hat{R}_t$. Using Eq. (4.16) and $\hat{\mathbf{R}}_v = (\mathbf{T}_2\mathbf{P}_2^{\mathrm{T}})^{\mathrm{T}}(\mathbf{T}_2\mathbf{P}_2^{\mathrm{T}})$ leads to:

$$\begin{aligned}\mathbf{w}_{\mathrm{SO-PLSR}} &= [\mathbf{P}_1\mathbf{a}_1\mathbf{a}_1^{\mathrm{T}}\mathbf{P}_1^{\mathrm{T}} + \mathbf{P}_2\mathbf{a}_2\mathbf{a}_2^{\mathrm{T}}\mathbf{P}_2^{\mathrm{T}} + \mathbf{P}_1\mathbf{a}_1\mathbf{a}_2^{\mathrm{T}}\mathbf{P}_2^{\mathrm{T}} +\\ &\quad \mathbf{P}_2\mathbf{a}_2\mathbf{a}_1^{\mathrm{T}}\mathbf{P}_1^{\mathrm{T}} + \mathbf{P}_2\mathbf{T}_2^{\mathrm{T}}\mathbf{T}_2\mathbf{P}_2^{\mathrm{T}}/\hat{R}_t]^{+}(\mathbf{P}_1\mathbf{a}_1 + \mathbf{P}_2\mathbf{a}_2)\,,\\ &= \left(\begin{bmatrix}\mathbf{P}_1 & \mathbf{P}_2\end{bmatrix}\begin{bmatrix}\mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_1\mathbf{a}_2^{\mathrm{T}}\\ \mathbf{a}_2\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_2\mathbf{a}_2^{\mathrm{T}} + \mathbf{C}\end{bmatrix}\begin{bmatrix}\mathbf{P}_1^{\mathrm{T}}\\ \mathbf{P}_2^{\mathrm{T}}\end{bmatrix}\right)^{+}(\mathbf{P}_1\mathbf{a}_1 + \mathbf{P}_2\mathbf{a}_2)\,,\\ &= \begin{bmatrix}\mathbf{P}_1 & \mathbf{P}_2\end{bmatrix}\begin{bmatrix}\mathbf{a}_1\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_1\mathbf{a}_2^{\mathrm{T}}\\ \mathbf{a}_2\mathbf{a}_1^{\mathrm{T}} & \mathbf{a}_2\mathbf{a}_2^{\mathrm{T}} + \mathbf{C}\end{bmatrix}^{+}\begin{bmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{bmatrix}\,. \end{aligned} \tag{4.20}$$

Using Lemma 2, it is straightforward to show that Eq. (4.20) reduces to $\mathbf{w}_{\mathrm{SO-PLSR}} \propto \mathbf{P}_1\mathbf{a}_1$. Hence, $\mathbf{t}_{\mathrm{PCA-PLSR}} \propto \mathbf{t}_{\mathrm{SO-PLSR}}$. This proves that the first scores and loading vectors from PCA-PLSR and SO-PLSR are equivalent up to a scaling factor. The proportionality can be replaced by equality if the loading vectors are normalized. After deflation, as defined in Eq. (4.10), the next scores and loading vectors are also equal. □

4.3.2 Proposition 4: Equivalence of various methods with $r_{\mathrm{PLSR}} = r_{\mathrm{PCR}}$

Proposition 4 *With $r_{\mathrm{PLSR}} = r_{\mathrm{PCR}}$, all methods IT-PCR, PCA-PLSR, NSO-PLSR and SO-PLSR lead to the same regression vector.*

Proof:

NSO-PLSR leads to the same regression vector as IT-PCR for $r_{\mathrm{PLSR}} = r_{\mathrm{PCR}}$ (for proof see [67]). Next, consider PCA-PLSR with $r_{\mathrm{PLSR}} = r_{\mathrm{PCR}}$. Since there is no dimensionality reduction, the PLSR step in PCA-PLSR is equivalent to least-squares, hence $\hat{\mathbf{b}}_{\mathrm{PCA-PLSR}} = \mathbf{P}_{\mathrm{IT-PCR}} \left(\tilde{\mathbf{X}}_c \mathbf{P}_{\mathrm{IT-PCR}}\right)^{+} \tilde{\mathbf{y}}_c = \hat{\mathbf{b}}_{\mathrm{IT-PCR}}$. Lastly, by Proposition 3, the equivalence of scores and loadings of SO-PLSR and PCA-PLSR implies that of the regression vectors, which completes the proof. □

Note that this proposition does not state the equivalence of scores and loadings (which is a stronger condition), but only the equivalence of predictions from IT-PCR, PCA-PLSR, NSO-PLSR and SO-PLSR.

4.4 Illustrative examples

Five examples are presented in this section. Simulated data to motivate the use of unlabeled data and to study the effect of drift are presented in Examples 1 and 2, respectively. The two propositions are illustrated using experimental data in Example 3. Finally, Examples 4 and 5 use experimental data to show the advantage of using unlabeled data at different X-noise levels.

4.4.1 First example (simulated data): Motivation to use unlabeled data

The use of unlabeled data together with labeled data is motivated with a toy example of a two-component mixture. The labeled, unlabeled and validation data are generated for the case of no drift, using Eq. (2.12), with the pure component spectra $\mathbf{S}$ as shown in Fig. 4.4. Fig. 4.5 shows schematically $\mathbf{Z}_c$ and $\mathbf{Z}_e$ for 6 different cases. The analyte of interest is the first mixture component, for which the y-values are available with measurement noise v_y, i.e. $\tilde{y} = z_1 + v_y$, where $v_y \sim \mathcal{N}(0, 0.1)$. X-noise is generated from $\mathcal{N}(0, 2)$. Spectra at $n_x = 100$ channels, measured for $n_c = 10$ mixture samples, are used for calibration using PCR and IT-PCR with 2 latent vectors. $\tilde{\mathbf{X}}_e$ from hundred samples ($n_e = 100$) is assumed to be available to correct the model. An independent test set $\{\tilde{\mathbf{X}}_p, \tilde{\mathbf{y}}_p\}$ from hundred samples ($n_p = 100$) is used for validation, with $\mathbf{Z}_p$ generated from the same probability distribution as $\mathbf{Z}_e$.

The following cases are shown in Fig. 4.5: (1) $\mathbf{Z}_c$ and $\mathbf{Z}_e$ are sampled from the same probability density functions, (2) and (3) $\mathbf{Z}_c$ and $\mathbf{Z}_e$ have different variances, (4) $\mathbf{Z}_c$ and $\mathbf{Z}_e$ have different means, (5) $\mathbf{Z}_c$ and $\mathbf{Z}_e$ have different covariance structures, and (6) $\mathbf{Z}_e$ is rank deficient (e.g. samples collected on-line during the run may be correlated due to stoichiometry).

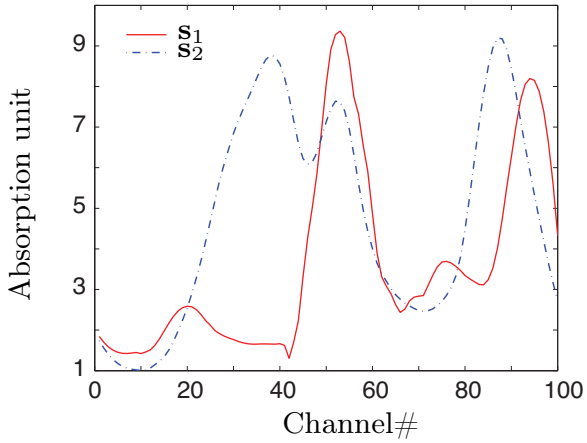


Fig. 4.4. First example: Pure component spectra used in simulation.

Since the measurements are free of drift, and noise-free $\mathbf{y}_p$ is used, RRMSEP_{e2} = RRMSEP_{e4}^{app} = 0. Fig. 4.6 shows RRMSEP, RRMSEP_{e1}, and RRMSEP_{e3} obtained from PCR and IT-PCR, and averaged over 1000 Monte Carlo simulations with different realizations of $\mathbf{Z}_c$, $\mathbf{Z}_e$ and measurement noise. For each of the six distributions, the error associated with the inaccurate modeling of subspace (RRMSEP_{e1}) is reduced through the use of unlabeled data due to the stabilization of the principal components (or better noise averaging), while the error due to noise in the prediction data (RRMSEP_{e3}) increases slightly due to the increase in the norm of the regression vector $||\hat{\mathbf{b}}||$. Overall, RRMSEP from IT-PCR is reduced because of larger contribution from the reduction in RRMSEP_{e1}. Fig. 4.7 shows that IT-PCR results in a smoother $\hat{\mathbf{b}}$ than that obtained with PCR, an indication of better noise averaging, and is closer to the NAS vector. The smoothening of $\hat{\mathbf{b}}$ can result in an increase or decrease of $||\hat{\mathbf{b}}||$. Note that, while RRMSEP_{e1} is always reduced for drift-free X-measurements, RRMSEP_{e3} may either increase or decrease.

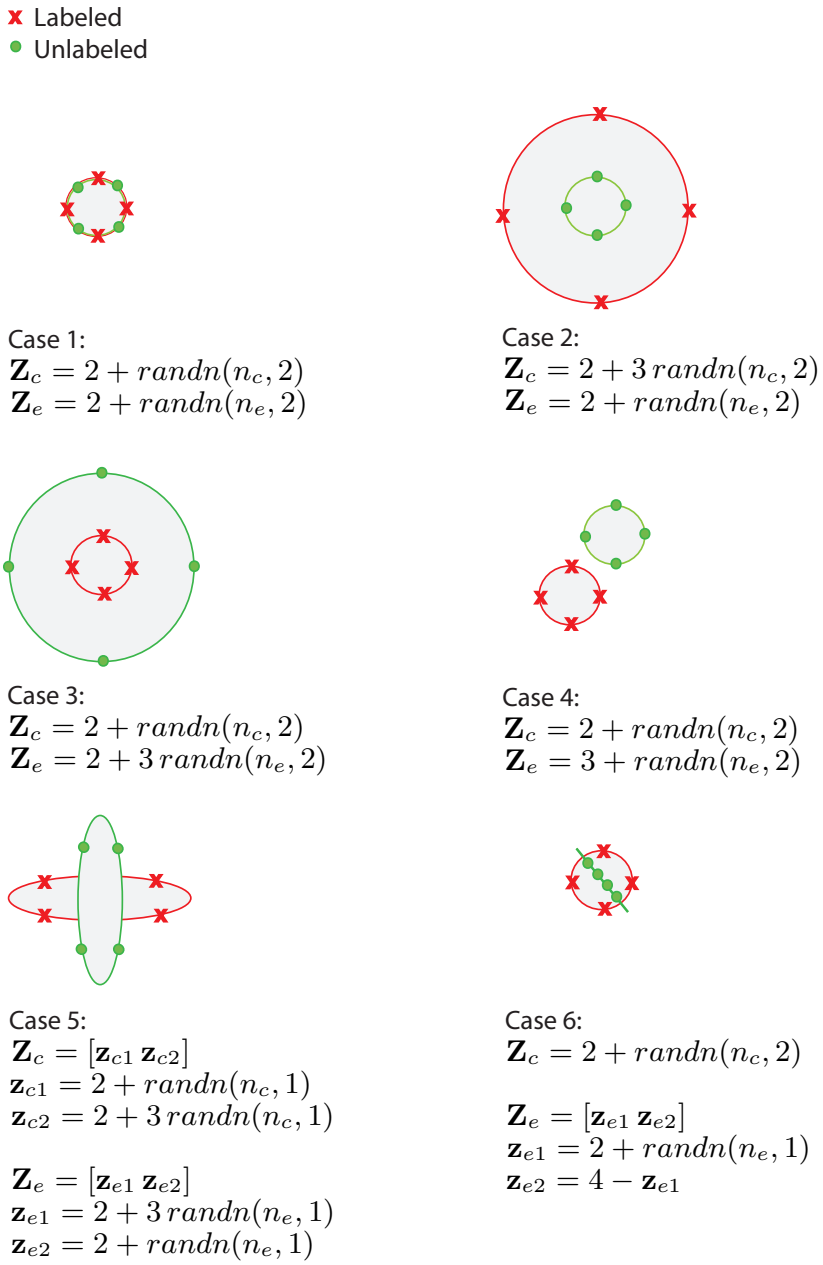


Fig. 4.5. First example: Schematic diagram of $\mathbf{Z}_c$ and $\mathbf{Z}_e$ (generated using MATLAB command $randn(g, m)$ that creates a $g \times m$ matrix whose samples are drawn from a normal distribution with zero mean and unit variance).

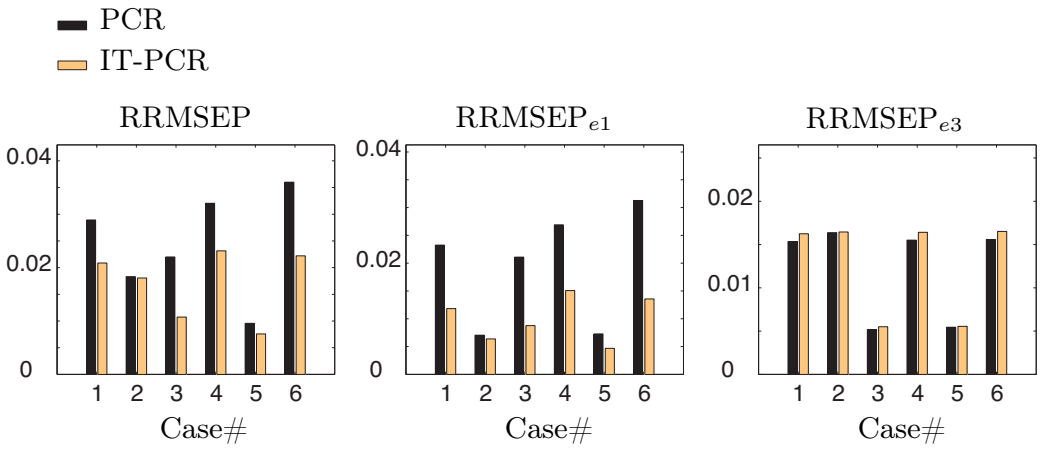


Fig. 4.6. First example: RRMSEP, RRMSEP$_{e1}$, and RRMSEP$_{e3}$ using standard (without unlabeled data) PCR and IT-PCR (with unlabeled data) for the six cases illustrated in Fig. 4.5.

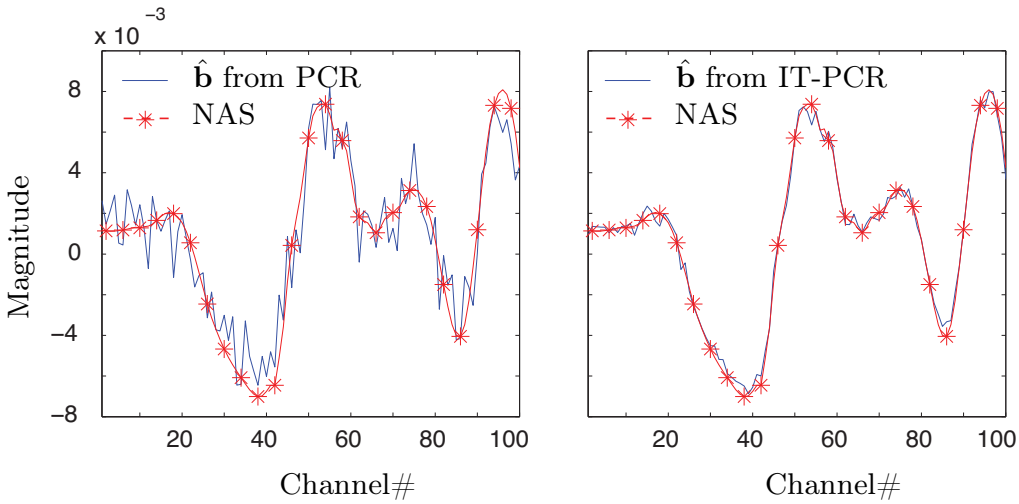


Fig. 4.7. First example: $\hat{\mathbf{b}}$ for the first analyte using standard (without unlabeled data) PCR and IT-PCR (with unlabeled data) for the first case illustrated in Fig. 4.5. IT-PCR results in a smoother $\hat{\mathbf{b}}$ than that obtained with PCR, and is closer to the NAS.

4.4.2 Second example (simulated data): Study of drift

The simulation example used in [67] illustrates the effect of drift in the unlabeled data on the three error terms individually. The description of the simulated data is repeated here for convenience. For a three-component mixture, the frequency spectrum in the range $0 < f \leq 100$ frequency units is obtained as:

$$\tilde{x}(f) = \frac{f_1\, f\, z_1}{\sqrt{(f_1^2 - f^2)^2 + (2\zeta_1 f_1 f)^2}} + \frac{f_2\, f\, z_2}{\sqrt{(f_2^2 - f^2)^2 + (2\zeta_2 f_2 f)^2}} + \frac{f_3\, f\, z_3}{\sqrt{(f_3^2 - f^2)^2 + (2\zeta_3 f_3 f)^2}} + v(f)\,, \tag{4.21}$$

with resonance frequencies $f_1 = 40$ fu, $f_2 = 50$ fu, $f_3 = 60$ fu, and the relative dampings $\zeta_1 = \zeta_2 = \zeta_3 = 0.05$. It is assumed that $z_1 \sim \mathcal{N}(3, 1)$, $z_2 \sim \mathcal{N}(3, 1)$, $z_3 \sim \mathcal{N}(3, \sqrt{0.5})$, and $v(f) \sim \mathcal{N}(0, 5)$. With $n_c = 40$, the $[40 \times 100]$ $\tilde{\mathbf{X}}_c$ is constructed as in Eq. (2.12), with the $[100 \times 3]$ matrix $\mathbf{S}$ defined from Eq. (4.21). With $n_e = 160$ and $n_p = 200$, the $[160 \times 100]$ $\tilde{\mathbf{X}}_e$ and the $[200 \times 100]$ $\tilde{\mathbf{X}}_p$ are constructed as in Eq. (2.19), using the same $\mathbf{S}$ and a randomly generated, smooth baseline $\mathbf{d}_p$. The baseline is assumed to be invariant over time, hence, $\mathbf{D}_p = \mathbf{1}_{n_p}\mathbf{d}_p^{\mathrm{T}}$. The analyte of interest is the second mixture component, for which the corresponding y-values are available with measurement noise v_y, i.e. $\tilde{y} = z_2 + v_y$, where $v_y \sim \mathcal{N}(0, 0.1)$.

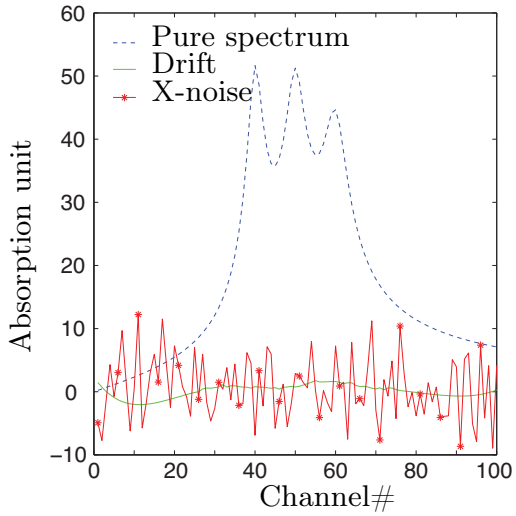


Fig. 4.8. Second example: X-noise, drift and pure spectrum (i.e. noise-free and drift-free) plotted separately.

Fig. 4.8 shows the three components of a spectrum belonging to the unlabeled data: the pure (noise-free and drift-free) spectrum, drift with l_2-norm 9.4, and X-noise. Fig. 4.9 shows RRMSEP as a function of the l_2-norm of the drift.

This study shows that PCA-PLSR performs well if the new data are drift-free. However, in the presence of drift, PCA-PLSR can lead to even larger prediction errors than by standard PLSR. Even a drift of small magnitude (compare the X-noise and drift magnitudes in Fig. 4.8) offsets the advantage gained from the stabilization of the principal component subspace.

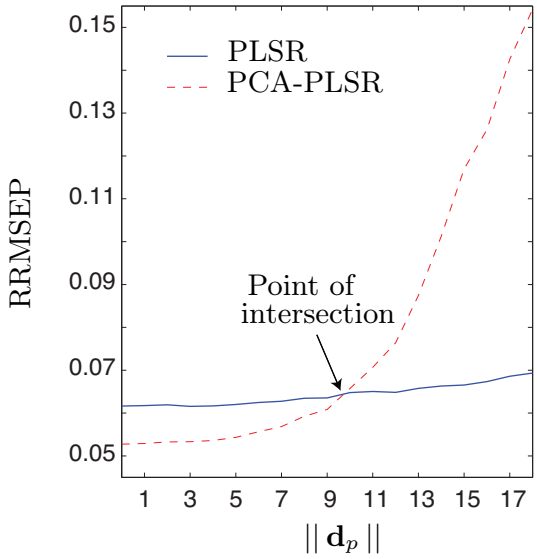


Fig. 4.9. Second example: RRMSEP averaged over 1000 Monte Carlo simulations.

Fig. 4.10 shows that each of the error terms $e1$, $e2$, and $e3$ increases with $||\mathbf{d}_p||$. RRMSEP_{e1} increases with the increase in drift that causes $\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{b}}$, $\mathbf{s}_2^{\mathrm{T}}\hat{\mathbf{b}}$, and $\mathbf{s}_3^{\mathrm{T}}\hat{\mathbf{b}}$ to deviate further away from the desired values 0, 1, and 0, respectively (see Fig. 4.10iv). Let θ denote the angle between $\hat{\mathbf{b}}$ and $\mathbf{d}_p$ vectors. The $|cos(\theta)|$ plot shows that $\hat{\mathbf{b}}$ is rotated towards $\mathbf{d}_p$ (see Fig. 4.10v), thereby increasing the inner product $\hat{\mathbf{b}}^{\mathrm{T}}\mathbf{d}_p$ and hence RRMSEP_{e2}. The increase in $||\hat{\mathbf{b}}||$ (see Fig. 4.10vi) results in an increase in RRMSEP_{e3}. In contrast to the simulation example in Section 4.4.1, at $||\mathbf{d}_p|| = 0$ (case of no drift), $||\hat{\mathbf{b}}||$ from PCA-PLSR is smaller than that from PLSR. As discussed in Section 4.4.1, the use of unlabeled data can result in an increase or decrease of RRMSEP_{e3}. However, RRMSEP_{e1} is reduced as expected for the case of no drift. Each of $\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{b}}$, $\mathbf{s}_2^{\mathrm{T}}\hat{\mathbf{b}}$, and $\mathbf{s}_3^{\mathrm{T}}\hat{\mathbf{b}}$ is closer to the desired value at $||\mathbf{d}_p|| = 0$. Similar conclusions can be drawn also for IT-PCR.

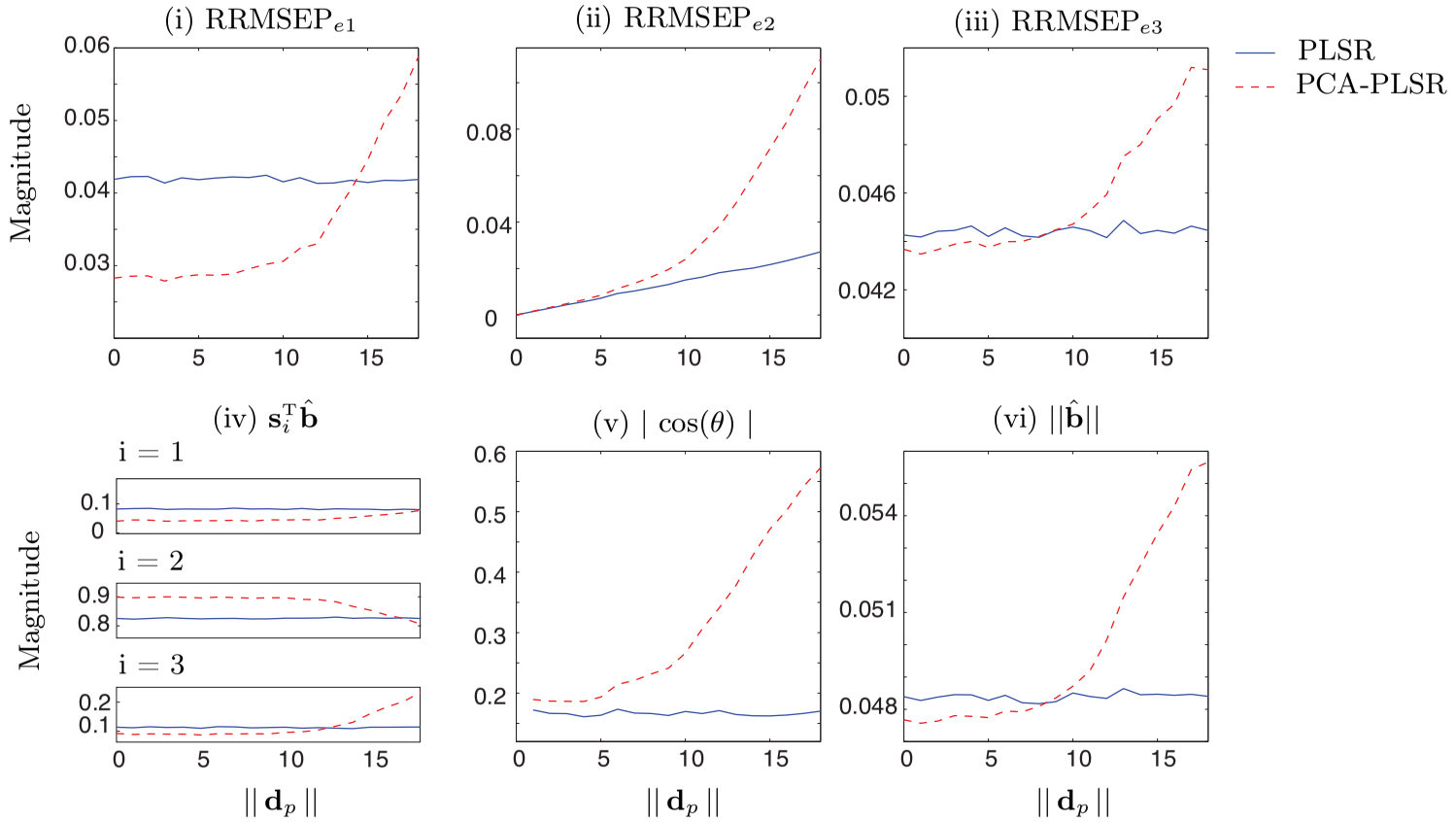


Fig. 4.10. Second example: Figures (i), (ii) and (iii) show the components of RRMSEP due to subspace modeling error, drift, and X-noise in prediction data, respectively. Figure (iv) shows that $\mathbf{s}_2^{\mathrm{T}}\hat{\mathbf{b}} \neq 1$, $\mathbf{s}_1^{\mathrm{T}}\hat{\mathbf{b}}$ and $\mathbf{s}_3^{\mathrm{T}}\hat{\mathbf{b}} \neq 0$, thereby leading to subspace modeling error. With the increase in $||\mathbf{d}_p||$, the deviation from 1 or 0 increases, thereby increasing RRMSEP$_{e1}$. Figure (v) shows that, with the increase in $||\mathbf{d}_p||$, $\hat{\mathbf{b}}$ rotates more towards $\mathbf{d}_p$, thereby increasing the inner product $\hat{\mathbf{b}}^{\mathrm{T}}\mathbf{d}_p$ and hence RRMSEP$_{e2}$. Figure (vi) shows that, with the increase in $||\mathbf{d}_p||$, $||\hat{\mathbf{b}}||$ also increases, thereby increasing RRMSEP$_{e3}$. Each plot is averaged over 1000 Monte Carlo simulations.

4.4.3 Third example (experimental data): Illustration of Propositions 3 and 4

This data set is from a designed experiment involving mixtures of three metal ions (Co(II), Cr(III), and Ni(II)) [70, 71]. The X-measurements are absorbance spectra recorded at 176 wavelengths over the range of 300-650 nm on a HP 8452 diode array spectrophotometer, while the concentration of Cobalt is used as the response variable. Five replicate spectra were obtained for each of the twenty-six mixtures using randomized blocks (i.e. 5 blocks of 26 mixtures, randomly ordered within each block, leading to a total of 130 measurements). To minimize the effects of instrumental drift, a reference spectrum was run prior to each new sample. As in [67], the data are auto-scaled and partitioned such that $n_c = 20$, $n_e = 84$ and $n_p = 26$. With r_{PCR} chosen as 5, the RRMSEP^{app} is shown in Table 4.1. Since PLSR uses more degrees of freedom per factor than PCR [72], often fewer factors are required in PLSR than in PCR [73], and so the study is limited to cases with $r_{\text{PLSR}} \leq 5$.

The following observations can be made: SO-PLSR and NSO-PLSR lead to slightly different RRMSEPs^{app} for certain values of r_{PLSR}, while, as expected, SO-PLSR and PCA-PLSR lead to equivalent RRMSEPs^{app} for all values of r_{PLSR}. Proposition 3 was verified numerically by confirming that the scores and loadings in PCA-PLSR and SO-PLSR were the same. The non-sequential nature of NSO-PLSR is illustrated in Fig. 4.11 by showing that the first scores and loading vectors are different for $r_{\text{PLSR}} = 1, \ldots, 5$. Although the scores and loadings appear to overlap, the zoomed region shows that small differences exist. Though not shown here, the second and higher scores and loading vectors too have only minor variations due to the non-sequential evaluation. Thus, NSO-PLSR and SO-PLSR model nearly the same space, thereby leading to nearly the same regression vectors. Hence, it can be seen in Table 4.1 that for $1 < r_{\text{PLSR}} < r_{\text{PCR}}$, SO-PLSR and NSO-PLSR differ by less than 0.05%. When $r_{\text{PLSR}} = 1$, NSO-PLSR is equivalent to SO-PLSR and hence the RRMSEPs^{app} are the same. For $r_{\text{PLSR}} = r_{\text{PCR}} = 5$, both NSO-PLSR and SO-PLSR become a similarity transform of IT-PCR, and hence the RRMSEPs^{app} are again equal (see the bold numerical values in Table 4.1). Proposition 4 is thus verified numerically.

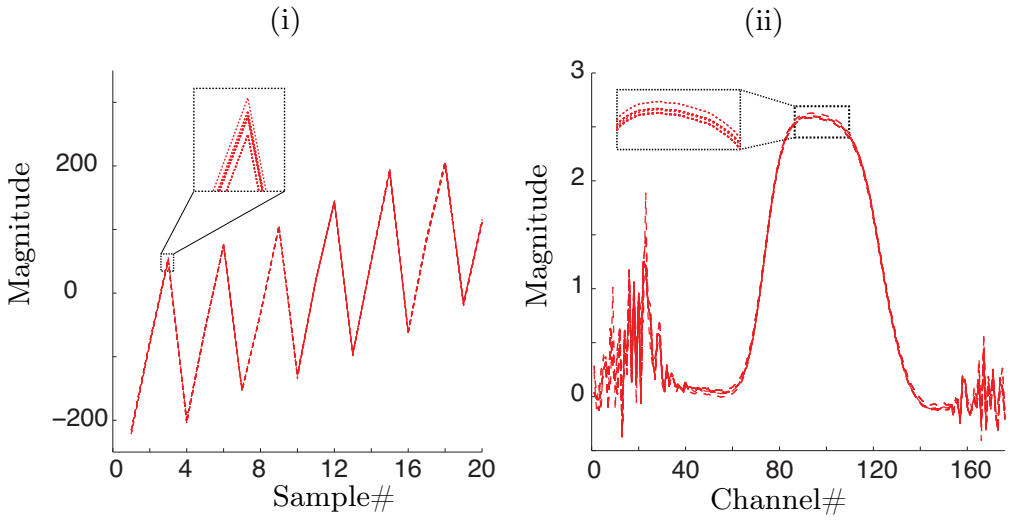


Fig. 4.11. Third example: (i) First scores vector, and (ii) first loading vector of NSO-PLSR for $r_{\text{PLSR}} = 1, \ldots, 5$. Although the vectors appear to overlap, the zoomed region shows that small differences exist since this method is non-sequential.

r_{PCR}	PCR	IT-PCR
1	47.992	49.090
2	17.788	18.859
3	3.566	3.346
4	3.694	2.586
5	3.694	**2.489**

r_{PLSR}	PLSR	NSO-PLSR	SO-PLSR or PCA-PLSR
1	42.146	**42.271**	**42.271**
2	5.332	4.030	4.039
3	3.429	2.393	2.355
4	3.497	2.592	2.594
5	3.402	**2.489**	**2.489**

Table 4.1. For the experimental metal ion data, %RRMSEP^{app} (i.e. RRMSEP^{app} $\times$ 100) obtained from standard (without unlabeled data) PCR/PLSR models and the different regression models with unlabeled data.

4.4.4 Fourth example (experimental data): Study of different X-noise levels

This data set is from the "Shootout" at the International Diffuse Reflectance Conference (IDRC 2002) involving 654 pharmaceutical tablets analyzed for assay value (the name of the active ingredient was not disclosed for proprietary reasons), tablet weight, and tablet hardness [74]. The X-measurements are NIR transmittance spectra recorded at 650 wavelengths over the range of 600-1898 nm on a Foss NIRSystems Multitab spectrometer, while the concentration of assay value is used as the response variable.

The reduction in RRMSEP^{app} due to stabilization of the latent subspace is illustrated at different SNR values by artificially adding noise to the X-measurements. Zero-mean Gaussian noise with standard deviation σ_v is added to achieve different noise levels $0 \leq \sigma_v \leq \sigma_x$, where σ_x is the average standard deviation in X-variables. The RRMSEP^{app} with PCA-PLSR and PLSR are computed as a function of σ_v/σ_x. For their comparison, the *oracle* values of meta-parameters (r_{PLSR} in PLSR, $\{r_{\text{PCR}}, r_{\text{PLSR}}\}$ in PCA-PLSR) are chosen, i.e. the calibration model is developed for all values of r_{PCR} and r_{PLSR} and those leading to the minimum RRMSEP^{app} are selected. The data are mean-centered and randomly partitioned such that $n_c = 100$, $n_e = 300$ and $n_p = 255$. At each X-noise level ($0 \leq \sigma_v \leq \sigma_x$), the RRMSEPs^{app} are averaged over 100 Monte Carlo runs, and in each run the data are randomly partitioned as labeled, unlabeled and prediction data.

The averaged RRMSEPs^{app} with PLSR and PCA-PLSR are shown in Fig. 4.12. As expected, the advantage gained from using unlabeled data is more significant when the X-measurements are noisy.

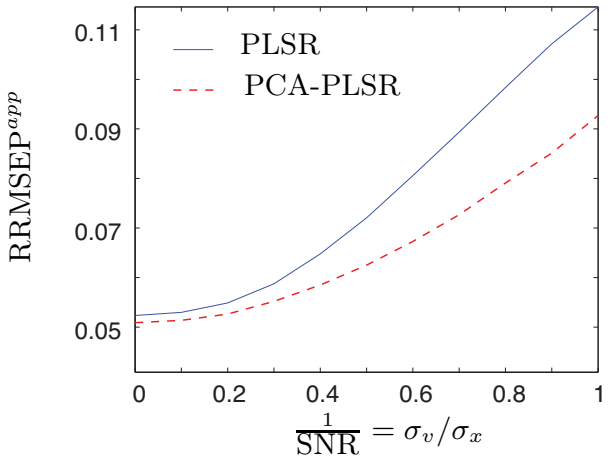


Fig. 4.12. Fourth example: RRMSEP^{app} averaged over 100 Monte Carlo runs, using standard (without unlabeled data) PLSR and PCA-PLSR (with unlabeled data) at different SNR values.

4.4.5 Fifth example (experimental data): Unlabeled data can replace labeled data

This data set is from a designed experiment involving mixtures of three alcohols (propanol, butanol, and pentanol) [75]. ^{1}H NMR spectra were recorded at 14000 shifts over the range of 3.85-0.65 ppm on a Bruker Avance Ultra Shield 400 spectrometer. The X-measurements are spectral intensities selected at 3000 shifts, while the concentration of propanol is used as the response variable. Each alcohol is varied on 21 concentration levels in increments of 5% from 0% to 100%, resulting in 231 mixtures.

In this example, it is shown that (sometimes freely available) unlabeled data help reduce the amount of (costly) labeled data required for a particular RRMSEP^{app}. For PCA-PLSR, the data are mean-centered and randomly partitioned such that $n_c = 8$, $n_e = 16$ and $n_p = 207$. The RRMSEP^{app} with PCA-PLSR is compared to that with PLSR built on $n_{c,a}$ labeled data, where $n_{c,a} = n_c + n_a$, and n_a represents additional labeled data required such that $\text{RRMSEP}^{app}_{\text{PLSR}} = \text{RRMSEP}^{app}_{\text{PCA}-\text{PLSR}}$.

As in the previous example, (i) zero-mean Guassian noise is added to the X-measurements, and (ii) oracle values of meta-parameters are chosen for model building. At each X-noise level ($0 \leq \sigma_v \leq \sigma_x$), the RRMSEPs^{app} are averaged over 100 Monte Carlo runs, and in each run the data are randomly partitioned as labeled, unlabeled and prediction data.

The value of n_a found from the Monte Carlo study is plotted as a function of σ_v/σ_x in Fig. 4.13. It can be seen that, when no artificial X-noise is added, 16 unlabeled samples are equivalent (in the sense that they improve RRMSEP^{app} by the same amount) to 1 additional labeled sample. However, at $\sigma_v/\sigma_x = 1$, 16 unlabeled samples are equivalent to 15 labeled samples! This example illustrates that, depending upon the expected SNR and the relative cost of obtaining labeled and unlabeled data, one may decide upon the number of (costly) labeled and (sometimes freely available) unlabeled data needed during calibration.

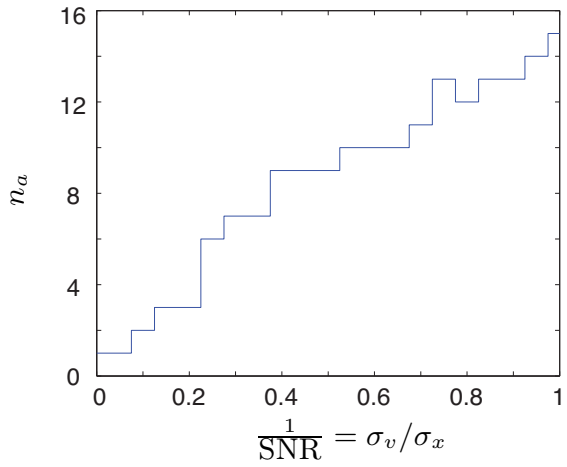


Fig. 4.13. Fifth example: Number of additional labeled data in PLSR at different SNR values such that, averaged over 100 Monte Carlo runs, $\text{RRMSEP}^{app}_{\text{PLSR}}$ = $\text{RRMSEP}^{app}_{\text{PCA}-\text{PLSR}}$ with $n_e = 16$.

4.5 Conclusions

Simulated and experimental data sets have been used to illustrate the usefulness and pitfalls of using unlabeled data. In the absence of drift, the use of unlabeled data helps stabilize the latent subspaces in the calibration step, thus leading to a lower RRMSEP. Hence, unlabeled data can replace labeled data to some extent and bring some economic benefit. However, in the presence of drift, the use of unlabeled data can result in an increase in prediction error compared to that obtained with a model based on labeled data alone. In prediction data, which by definition qualifies as unlabeled data, the presence of drift can be checked using the Q-statistic and a properly defined threshold. If drift is present, drift correction methods should be applied before using unlabeled data (see Chapter 3). This chapter has also discussed the equivalence of different methods using unlabeled data in PCR and PLSR. The only difference between NSO-PLSR and PCA-PLSR (or SO-PLSR) lies in the non-sequential estimation of factors in the NSO-PLSR and the sequential estimation in PCA-PLSR. Furthermore, it was shown analytically that, with $r_{\text{PLSR}} = r_{\text{PCR}}$, all methods IT-PCR, PCA-PLSR, NSO-PLSR and SO-PLSR lead to the same regression vector. For the examples considered, the difference in the RRMSEP with PCA-PLSR and NSO-PLSR is less than 0.05%. PCA-PLSR may be preferred over OF-based methods due to its simplicity.

5

Data reconciliation based on balance equations

This chapter studies data reconciliation (to reduce the overall RRMSEP) based on additional information in the form of balance equations.

5.1 Overview

Calibration models to predict the concentrations of various analytes from spectroscopic measurements are typically developed off-line. Such calibration models do not take into account certain relationships that exist amongst the various analytes e.g. due to the presence of chemical reactions. Additional on-line measurements (from e.g. flow rate devices and gas analyzers) and prevailing mass and elemental balance equations, along with the criteria of monotonicity, smoothness and non-negativity of concentrations, can be used to adjust and improve the predictions [76]. Data reconciliation (DR) is the procedure of optimally adjusting process measurements to obtain more accurate estimates, which are consistent with the balance equations and other constraints, to detect gross errors and estimate unmeasured variables. DR has been traditionally applied to chemical processes involving flow circuits, where model equation and mass balance equations are readily definable [77]. In this chapter, the use of DR for spectroscopic calibration is discussed. The following simplifying assumptions are made:

A1: The concentrations of all n_s analytes are predicted from the X-measurements, i.e. $n_k = n_s$, and

A2: the true concentrations $\mathbf{z}_p$ of the prediction set follow n_g balance equations:

$$\mathbf{G}\,\mathbf{z}_p = \mathbf{c}\,, \tag{5.1}$$

where $\mathbf{G}$ is an $[n_g \times n_s]$ matrix and $\mathbf{c}$ an n_g-dimensional vector.

The prediction error for the i^{th} analyte can be expressed according to Eq. (2.22),

$$\epsilon_{p,i} := \quad y_{p,i} - \hat{y}_{p,i} \quad = \quad \underbrace{(y_{p,i} - \mathbf{z}_p^{\mathrm{T}}\,\mathbf{S}^{\mathrm{T}}\hat{\mathbf{b}}_i - \mathbf{d}_{p||c}^{\mathrm{T}}\hat{\mathbf{b}}_i)}_{e1} - \underbrace{\mathbf{d}_{p\perp c}^{\mathrm{T}}\hat{\mathbf{b}}_i}_{e2} - \underbrace{\mathbf{v}_{x,p}^{\mathrm{T}}\,\hat{\mathbf{b}}_i}_{e3}\,. \tag{5.2}$$

The predicted values of the n_s analyte concentrations are given by:

$$\hat{\mathbf{z}}_p = \hat{\mathbf{B}}^{\mathrm{T}}\,\tilde{\mathbf{x}}_p = \mathbf{z}_p + \boldsymbol{\epsilon}_p\,, \tag{5.3}$$

where $\hat{\mathbf{B}} = [\hat{\mathbf{b}}_1 \ldots \hat{\mathbf{b}}_{n_s}]$ is the $[n_x \times n_s]$ matrix of regression vectors, and $\boldsymbol{\epsilon}_p = [\epsilon_{p,1}, \ldots, \epsilon_{p,n_s}]$ the n_s-dimensional prediction error vector with the dispersion matrix $\mathcal{E}[\boldsymbol{\epsilon}_p\,\boldsymbol{\epsilon}_p^{\mathrm{T}}] = \boldsymbol{\Psi}$ of size $[n_s \times n_s]$. Let the $[n_s \times n_s]$ empirical dispersion matrix $\boldsymbol{\Psi}_e$ of $\boldsymbol{\epsilon}_p$ be defined by:

$$\boldsymbol{\Psi}_e := \frac{1}{n_p}\mathbf{F}\mathbf{F}^{\mathrm{T}}\,, \tag{5.4}$$

where $\mathbf{F} = [\boldsymbol{\epsilon}_p(1) \ldots \boldsymbol{\epsilon}_p(n_p)]$ is an $[n_s \times n_p]$ matrix composed of prediction errors from n_p measurements. Typically, $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}_e$ are unknown and approximated by an $[n_s \times n_s]$ matrix $\mathbf{H}$. DR is formulated as the following optimization problem with a weighted l_2-norm cost function, and balance equations as linear equality constraints:

$$\hat{\boldsymbol{\epsilon}}_p = argmin\left[\boldsymbol{\epsilon}^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}\right] \quad \text{such that} \quad \mathbf{G}(\hat{\mathbf{z}}_p - \boldsymbol{\epsilon}) = \mathbf{c}\,, \tag{5.5}$$

which has the closed-form solution[1] [77]:

[1] Note that DR is invariant to multiplication of $\mathbf{H}$ by any scalar

$$\begin{aligned}\hat{\boldsymbol{\epsilon}}_p = &\quad \mathbf{H}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}})^{-1}(\mathbf{G}\hat{\mathbf{z}}_p - \mathbf{c}) = \mathbf{H}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G}\,\boldsymbol{\epsilon}_p \\ = &\quad \mathbf{M}_{\mathrm{H}}\,\boldsymbol{\epsilon}_p\,,\end{aligned} \tag{5.6}$$

where $\mathbf{M}_{\mathrm{H}} = \mathbf{H}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G}$ is an $[n_s \times n_s]$ square matrix. Since $\mathbf{M}_{\mathrm{H}}$ is an idempotent matrix (i.e. $\mathbf{M}_{\mathrm{H}}\mathbf{M}_{\mathrm{H}} = \mathbf{M}_{\mathrm{H}}$), it is a projection matrix. The reconciled predictions $\hat{\mathbf{z}}_r$ are:

$$\hat{\mathbf{z}}_r = \hat{\mathbf{z}}_p - \hat{\boldsymbol{\epsilon}}_p\,. \tag{5.7}$$

The residual error after DR, $\boldsymbol{\epsilon}_r := \hat{\mathbf{z}}_r - \mathbf{z}_p$ can be expressed as:

$$\boldsymbol{\epsilon}_r = \hat{\mathbf{z}}_p - \hat{\boldsymbol{\epsilon}}_p - \mathbf{z}_p = \boldsymbol{\epsilon}_p - \hat{\boldsymbol{\epsilon}}_p\,. \tag{5.8}$$

In spectroscopic calibration, estimating $\boldsymbol{\Psi}$ (or $\boldsymbol{\Psi}_e$) is a challenging task. In the absence of prior information, a naive assumption is to use $\mathbf{H} = \mathbf{I}_{n_s}$, leading to $\mathbf{M}_{\mathrm{I}} = \mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G} = \mathbf{G}^{+}\mathbf{G}$, i.e. orthogonal projection.

Consider the following toy example: $\mathbf{z}_p = \left[\begin{smallmatrix} y_{p,1} \\ y_{p,2} \end{smallmatrix}\right]$ where the variables $y_{p,1}$ and $y_{p,2}$ satisfy the balance equation $y_{p,1} + y_{p,2} = 1$ with $\mathbf{G} = [1 \quad 1]$ and $c = 1$ (see Fig. 5.1i). With $\mathbf{H} = \mathbf{I}_2 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, the reconciled solution $\hat{\mathbf{z}}_r$ is the nearest point (minimum l_2-norm) on the line that satisfies the balance equation, i.e. $\hat{\mathbf{z}}_p$ is orthogonally projected onto the line $y_{p,1} + y_{p,2} = 1$. Note that orthogonal projection leads to $||\mathbf{z}_p - \hat{\mathbf{z}}_r|| \leq ||\mathbf{z}_p - \hat{\mathbf{z}}_p||$ or $||\boldsymbol{\epsilon}_r|| \leq ||\boldsymbol{\epsilon}_p||$. However, this does not imply that both variables $y_{p,1}$ and $y_{p,2}$ have lower prediction error after DR. In this example, $|\hat{y}_{r,1} - y_{p,1}| < |\hat{y}_{p,1} - y_{p,1}|$ (improved prediction of y_1) but $|\hat{y}_{r,2} - y_{p,2}| > |\hat{y}_{p,2} - y_{p,2}|$ (worsened prediction of y_2).

If the standard deviation of errors in $y_{p,1}$ and $y_{p,2}$ were known, this prior knowledge could be used for oblique projection using $\mathbf{H} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & h_{22} \end{smallmatrix}\right]$. For example, if the variance of the $y_{p,1}$ prediction is 5 times larger than that of $y_{p,2}$, $h_{22} = 1/5$, thereby leading to an oblique projection that adjusts more $\hat{y}_{p,1}$ as opposed to $\hat{y}_{p,2}$ in order to satisfy the constraint (see Fig. 5.1ii). Using the uncertainty in y_2 relative to y_1, oblique projection will produce a lower residual error, i.e. $\mathcal{E}[||\mathbf{z}_p - \hat{\mathbf{z}}_r||_{\text{oblique}}] \leq ||\mathcal{E}[\mathbf{z}_p - \hat{\mathbf{z}}_r||_{\text{orthogonal}}]$. This example assumed a diagonal matrix $\mathbf{H}$, with larger entries chosen for more accurate predictions. This is a common assumption in the literature [76, 78–80]. However, such a choice of $\mathbf{H}$ ignores the correlation between the prediction errors of different

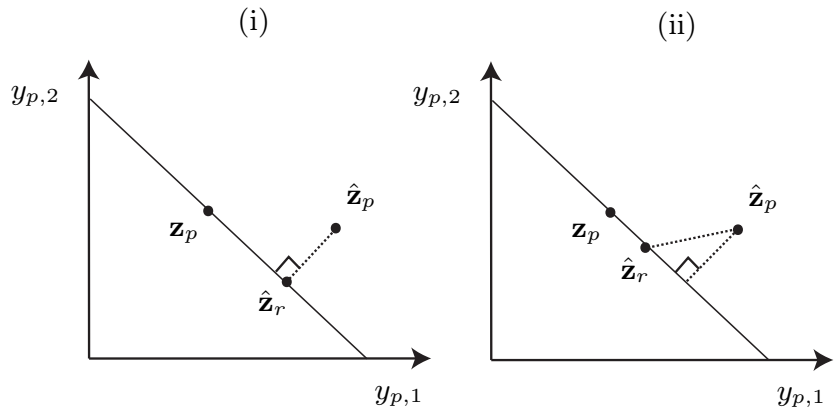


Fig. 5.1. (i) Orthogonal projection, (ii) oblique projection.

analytes. In the case of spectroscopic measurements, since the same X-noise $\mathbf{v}_{x,p}$ propagates to the prediction errors for different analytes, the prediction errors of the various analytes are indeed correlated.

The chapter is organized as follows. Section 5.2 motivates DR through two properties of the error after DR. The properties are illustrated with simulation and experimental studies in Section 5.3, and Section 5.4 concludes the chapter.

5.2 Features of DR

This section investigates two properties of the reconciled error $\boldsymbol{\epsilon}_r$ and discusses the different cases of $\mathbf{H}$ used in this work.

5.2.1 Proposition 5: Reduced overall error

The following Proposition says that the weighted l_2-norm of the prediction error after DR is smaller than that of prediction error before DR, regardless of the choice of $\mathbf{H}$.

Proposition 5 $\boldsymbol{\epsilon}_r^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_r \leq \boldsymbol{\epsilon}_p^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_p \ \forall \mathbf{H}$.

Proof: Using Eqs. (5.6) and (5.8),

$$\begin{aligned} \boldsymbol{\epsilon}_r^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_r = & \quad (\boldsymbol{\epsilon}_p - \hat{\boldsymbol{\epsilon}}_p)^{\mathrm{T}}\mathbf{H}^{-1}(\boldsymbol{\epsilon}_p - \hat{\boldsymbol{\epsilon}}_p) \\ = & \quad \boldsymbol{\epsilon}_p^{\mathrm{T}}(\mathbf{I}_{n_s} - \mathbf{M}_{\mathrm{H}})^{\mathrm{T}}\mathbf{H}^{-1}(\mathbf{I}_{n_s} - \mathbf{M}_{\mathrm{H}})\boldsymbol{\epsilon}_p \,. \end{aligned} \tag{5.9}$$

Substituting for $\mathbf{M}_{\mathrm{H}}$,

$$\begin{aligned} \boldsymbol{\epsilon}_r^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_r = & \quad \boldsymbol{\epsilon}_p^{\mathrm{T}}(\mathbf{H}^{-1} - \mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G})\boldsymbol{\epsilon}_p \\ = & \quad \boldsymbol{\epsilon}_p^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_p - \boldsymbol{\epsilon}_p^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G}\boldsymbol{\epsilon}_p \,. \end{aligned} \tag{5.10}$$

Since $\mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G}$ is positive semi-definite (PSD), $\boldsymbol{\epsilon}_p^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G}\boldsymbol{\epsilon}_p \geq 0$, thereby leading to $\boldsymbol{\epsilon}_r^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_r \leq \boldsymbol{\epsilon}_p^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_p$. □

5.2.2 Proposition 6: Reduced error for each analyte

The following Proposition says that with $\mathbf{H} = \boldsymbol{\Psi}$, RRMSEP for each analyte is reduced in expectation sense. The proof can be found in [77], and is restated here for convenience.

Proposition 6 $\mathbf{H} = \boldsymbol{\Psi}$ *leads to reduced expected RRMSEP for each analyte, i.e.* $\mathcal{E}[\epsilon_{r,i}^2] \leq \mathcal{E}[\epsilon_{p,i}^2]$, $\forall\ i = 1, \ldots, n_s$, *where* $\epsilon_{r,i}$ *and* $\epsilon_{p,i}$ *are the* i^{th} *elements of* $\boldsymbol{\epsilon}_r$ *and* $\boldsymbol{\epsilon}_p$, *respectively.*

Proof:

Using Eqs. (5.6) and (5.8),

$$\begin{aligned} \mathcal{E}[\boldsymbol{\epsilon}_r\boldsymbol{\epsilon}_r^{\mathrm{T}}] - \mathcal{E}[\boldsymbol{\epsilon}_p\boldsymbol{\epsilon}_p^{\mathrm{T}}] = & \quad \mathcal{E}[(\boldsymbol{\epsilon}_p - \mathbf{M}_{\mathrm{H}}\boldsymbol{\epsilon}_p)(\boldsymbol{\epsilon}_p - \mathbf{M}_{\mathrm{H}}\boldsymbol{\epsilon}_p)^{\mathrm{T}}] - \boldsymbol{\Psi} \\ = & \quad (\mathbf{I}_{n_s} - \mathbf{M}_{\mathrm{H}})\,\boldsymbol{\Psi}\,(\mathbf{I}_{n_s} - \mathbf{M}_{\mathrm{H}}^{\mathrm{T}}) - \boldsymbol{\Psi}\,. \end{aligned} \tag{5.11}$$

Substituting for $\mathbf{M}_{\mathrm{H}}$ when $\mathbf{H} = \boldsymbol{\Psi}$ leads to,

$$\mathcal{E}[\boldsymbol{\epsilon}_r\boldsymbol{\epsilon}_r^{\mathrm{T}}] - \mathcal{E}[\boldsymbol{\epsilon}_p\boldsymbol{\epsilon}_p^{\mathrm{T}}] = -\boldsymbol{\Psi}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\boldsymbol{\Psi}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G}\boldsymbol{\Psi}\,. \tag{5.12}$$

Since $\mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{H}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G}$ is PSD, diagonal elements of $\boldsymbol{\Psi}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\boldsymbol{\Psi}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{G}\boldsymbol{\Psi}$ are ≥ 0, and the diagonal elements of $\left(\mathcal{E}[\boldsymbol{\epsilon}_r\boldsymbol{\epsilon}_r^{\mathrm{T}}] - \mathcal{E}[\boldsymbol{\epsilon}_p\boldsymbol{\epsilon}_p^{\mathrm{T}}]\right)$ are ≤ 0. In other words, $\mathcal{E}[\epsilon_{r,i}^2] \leq \mathcal{E}[\epsilon_{p,i}^2], \quad \forall \quad i = 1, \ldots, n_s$. □

Replacing the expectation $\mathcal{E}[\boldsymbol{\epsilon}_r\boldsymbol{\epsilon}_r^{\mathrm{T}}] - \mathcal{E}[\boldsymbol{\epsilon}_p\boldsymbol{\epsilon}_p^{\mathrm{T}}]$, by empirical average $\sum_{j=1}^{n_p}\boldsymbol{\epsilon}_r(j)\boldsymbol{\epsilon}_r(j)^{\mathrm{T}} - \sum_{j=1}^{n_p}\boldsymbol{\epsilon}_p(j)\boldsymbol{\epsilon}_p(j)^{\mathrm{T}}$, leads to the following corollary.

Corollary of Proposition 6: $\mathbf{H} = \boldsymbol{\Psi}_e$ leads to decreased RRMSEP for each analyte, i.e. $\sum_{j=1}^{n_p} \epsilon_{r,i}(j)^2 \leq \sum_{j=1}^{n_p} \epsilon_{p,i}(j)^2, \quad \forall\, i = 1, \ldots, n_s$, where the summation is over n_p prediction samples.

5.2.3 Cases of H considered

The four cases of $\mathbf{H}$ studied in this work are described below and summarized in Table 5.1. This list of $\mathbf{H}$ is exemplary, not comprehensive.

(i) $\mathbf{H} = \mathbf{I}_{n_s}$: In the absence of prior knowledge, a naive assumption is to use the identity matrix, leading to orthogonal projection.

(ii) $\mathbf{H} = \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$: Assuming that the bias in the prediction errors (due to error terms $e1$ and $e2$) is negligible compared to the variance (due to error term $e3$), $\boldsymbol{\epsilon}_p \approx \hat{\mathbf{B}}^{\mathrm{T}}\,\mathbf{v}_{x,p}$. Hence $\mathcal{E}[\boldsymbol{\epsilon}_p\,\boldsymbol{\epsilon}_p^{\mathrm{T}}] \approx \mathcal{E}[\hat{\mathbf{B}}^{\mathrm{T}}\,\mathbf{v}_{x,p}(\hat{\mathbf{B}}^{\mathrm{T}}\,\mathbf{v}_{x,p})^{\mathrm{T}}] = \hat{\mathbf{B}}^{\mathrm{T}}\mathcal{E}[\mathbf{v}_{x,p}\mathbf{v}_{x,p}^{\mathrm{T}}]\hat{\mathbf{B}}$. Furthermore, assuming that the elements of $\mathbf{v}_{x,p}$ are iid from $\mathcal{N}(0,\sigma)$, $\mathcal{E}[\mathbf{v}_{x,p}\mathbf{v}_{x,p}^{\mathrm{T}}] = \sigma^2\mathbf{I}_{n_x}$. Hence, $\mathbf{H} = \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$ is one candidate approximation of $\boldsymbol{\Psi}$.

(iii) $\mathbf{H} = \mathbf{H}_{\mathrm{ASTM}}$: Assuming the off-diagonal elements of $\boldsymbol{\Psi}$ are negligible compared to the diagonal elements, the prediction confidence intervals of all analytes can be chosen as diagonal elements of $\mathbf{H}$. Several methods exist that approximate the confidence intervals [81–83]. We use the American Society for Testing and Materials (ASTM) standard [84], which allows to choose a sample specific $\mathbf{H}$. For the k^{th} prediction,

$$\begin{aligned} \mathbf{H}_{\mathrm{ASTM}}(k) &:= \quad diag\left(\sqrt{m_1\,(1+h_1(k))},\,\ldots\,,\sqrt{m_{n_s}\,(1+h_{n_s}(k))}\right) \\ m_i &:= \quad \sum_{j=1}^{j=n_c} \left(\tilde{y}_{c,i}(j) - \hat{y}_{c,i}(j)\right)^2/(n_c - df)\,, \end{aligned} \tag{5.13}$$

where $h_i(k)$ is the leverage of the i^{th} analyte for the k^{th} prediction [15], $y_{c,i}(j)$ the j^{th} calibration concentration for the i^{th} analyte, and df the degrees of freedom. In PCR, df is equal to the number of factors retained. Though PLSR uses more than one degree of freedom per factor [72], df is commonly approximated as the number of factors retained.

Furthermore, the loading space of PCR calibration models of n_s analytes are identical for the same number of factors retained. In this case, using $h_1(k) =$

$h_2(k) = \cdots = h_{n_s}(k)$, and the fact that DR is invariant to $\mathbf{H}(k)$ multiplied by a scalar, $\mathbf{H}(k)$ can be simplified to $\mathbf{H}(k) = \mathbf{H} = diag(\sqrt{m_1}, \ldots, \sqrt{m_{n_s}})$. In contrast to PCR, the loading space of the PLSR calibration model is different for each analyte.

(iv) $\mathbf{H} = \boldsymbol{\Psi}_e$: As shown in Proposition 6, $\mathbf{H} = \boldsymbol{\Psi}_e$ guarantees reduction in RRMSEP for each analyte.

Case of $\mathbf{H}$	Features
$\mathbf{I}_{n_s}$	orthogonal projection; naive
$\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$	oblique projection; ignores the bias
$\mathbf{H}_{\mathrm{ASTM}}$	oblique projection; ignores off-diagonal elements
$\boldsymbol{\Psi}_e$	oblique projection; reduced RRMSEP for each analyte

Table 5.1. Features of the different choices of $\mathbf{H}$.

5.3 Illustrative examples

Two examples are presented in this section. Example 1 compares RRMSEP before and after DR using simulated data. DR on experimental data is studied in Example 2.

5.3.1 First example (simulated data): Use of material balance equation

Generation of simulated data: DR is illustrated via spectroscopic measurements from a simulated isothermal constant-density batch reactor involving $n_s = 4$ absorbing analytes and 2 chemical reactions (example taken from [85]). Reactant A_1 is converted to the desired product A_4 following an auto-catalyzed two-step reaction:

$$2\,A_1 \xrightarrow{\kappa_1} A_2 \xrightarrow{\kappa_2} A_3 + 2\,A_4\,. \tag{5.14}$$

All four analytes are assumed to absorb. For calibration, $n_c = 30$ mixture samples are randomly generated ($\mathrm{rank}\,(\mathbf{Z}_c) = n_s$) within the operational concentration ranges. The spectroscopic measurements $\mathbf{X}_c$ are generated according to Eq. (2.12) using n_x-dimensional pure-component spectra ($n_x = 101$;

see Fig. 5.2i). Measurement noise with elements of $\mathbf{v}_x$ from $\mathcal{N}(0, 0.03)$, and $v_y \sim \mathcal{N}(0, 0.05)$, is added to the X and y-measurements. Three calibration models (NASR, PCR, and PLSR) are built for each analyte. For each analyte, four factors are chosen in its PCR and PLSR models. On-line spectroscopic

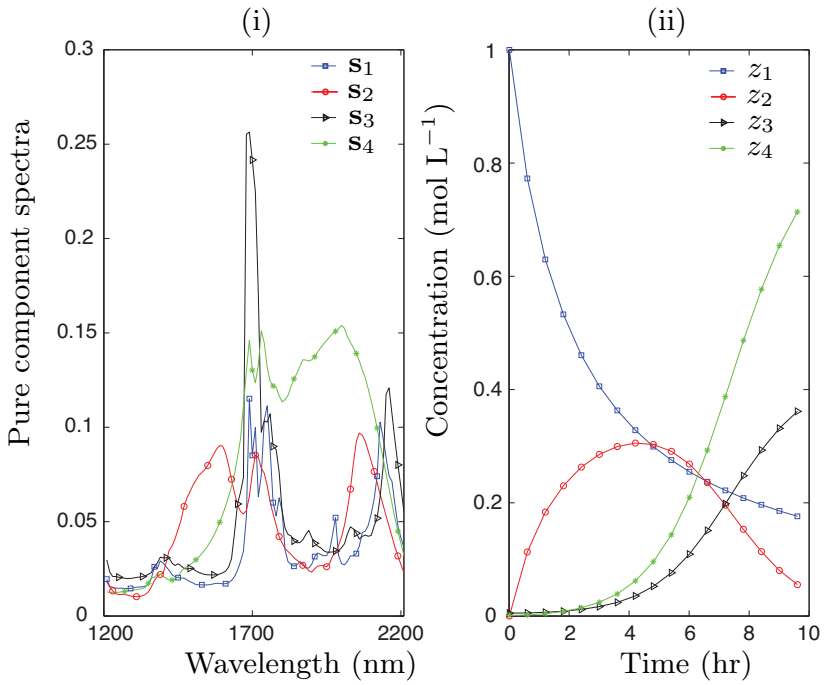


Fig. 5.2. Example 1: (i) Pure–component spectra, and (ii) concentration profiles of the four species, used for generating the simulated data.

measurements $\mathbf{x}_p$ are generated according to Eq. (2.12), where the concentrations are simulated from the following mole balances (see also Fig. 5.2ii):

$$\begin{aligned} \frac{d\,z_1}{d\,t} &= -2\kappa_1\, z_1^2 \\ \frac{d\,z_2}{d\,t} &= \kappa_1\, z_1^2 - \kappa_2\, z_2\, z_3\,, \end{aligned} \tag{5.15}$$

where z_i is the molar concentration of i^{th} analyte, and the numerical values of the rate constants are $\kappa_1 = 0.245\, L\ \text{mol}^{-1}\text{h}^{-1}$ and $\kappa_2 = 2.133\, L\ \text{mol}^{-1}\ \text{h}^{-1}$. Based on the null space of the stoichiometric matrix $\begin{bmatrix} -2 & 1 & 0 & 0 \\ 0 & -1 & 1 & 2 \end{bmatrix}$, the following $n_g = 2$ reaction-invariant relationships can be formulated:

$$\begin{bmatrix} 0.5 & 1 & 1 & 0 \\ 0.5 & 1 & 0 & 0.5 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} z_1(0) \\ z_2(0) \\ z_3(0) \\ z_4(0) \end{bmatrix}, \tag{5.16}$$

where $z_i(0)$ $(i = 1, \ldots, 4)$ refers to the initial analyte concentrations at the start of the reaction.

DR is illustrated for the three calibration models (NASR, PCR, and PLSR) and the four cases of $\mathbf{H}$ $(= \mathbf{I}_{n_s}$, $\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$, $\mathbf{H}_{\mathrm{ASTM}}$, and $\boldsymbol{\Psi}_e)$. Note that $\mathbf{H}_{\mathrm{ASTM}}$ is applicable for PCR and PLSR, but not for NASR.

Discussion: For one realization of the noise, the plots of $100\left(\frac{\boldsymbol{\epsilon}_r^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_r}{\boldsymbol{\epsilon}_p^{\mathrm{T}}\mathbf{H}^{-1}\boldsymbol{\epsilon}_p} - 1\right)$ are shown in Fig. 5.3. Negative values indicate a reduction in the weighted l_2-norm. Proposition 5 is satisfied for NASR, PCR, and PLSR for all the four cases of $\mathbf{H}$. For NASR, the difference in weighted error norm is randomly distributed because the prediction errors in consecutive samples are independent. In contrast to NASR, the difference in weighted error norm from PCR and PLSR is not random iid. This is because the bias in prediction error, resulting from the error term $e1$, depends upon the smoothly varying analyte concentrations. Hence, in consecutive samples, the prediction errors are highly correlated.

The RRMSEP of the four analytes, averaged over 100 Monte Carlo simulations for the three calibration models, and different cases of $\mathbf{H}$, are shown in Fig. 5.4. With $\mathbf{H} = \boldsymbol{\Psi}_e$ and $\mathbf{H} = \mathbf{H}_{\mathrm{ASTM}}$, RRMSEP is reduced for NASR, PCR, and PLSR, and for all four analytes. While this was expected for $\mathbf{H} = \boldsymbol{\Psi}_e$ (see Proposition 6), it does not hold in general for $\mathbf{H} = \mathbf{H}_{\mathrm{ASTM}}$. With $\mathbf{H} = \mathbf{I}_{n_s}$ and $\mathbf{H} = \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$, RRMSEP is not reduced for all combinations of calibration models and analytes. Unfortunately, without knowledge of $\boldsymbol{\Psi}_e$, one cannot determine in advance the predictions of which analytes will be worsened.

Note that, in the case of NASR, $\mathbf{H} = \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$ leads to approximately same RRMSEP as obtained with $\mathbf{H} = \boldsymbol{\Psi}_e$ since $\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}} \approx \boldsymbol{\Psi}_e$. In contrast, due to the subspace modeling errors in PCR and PLSR, $\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$ does not approximate $\boldsymbol{\Psi}_e$ well, thereby leading to different RRMSEP. In this example, the RRMSEP from PCR and PLSR, with $\mathbf{H} = \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$, are worse than even those obtained with $\mathbf{H} = \mathbf{I}_{n_s}$.

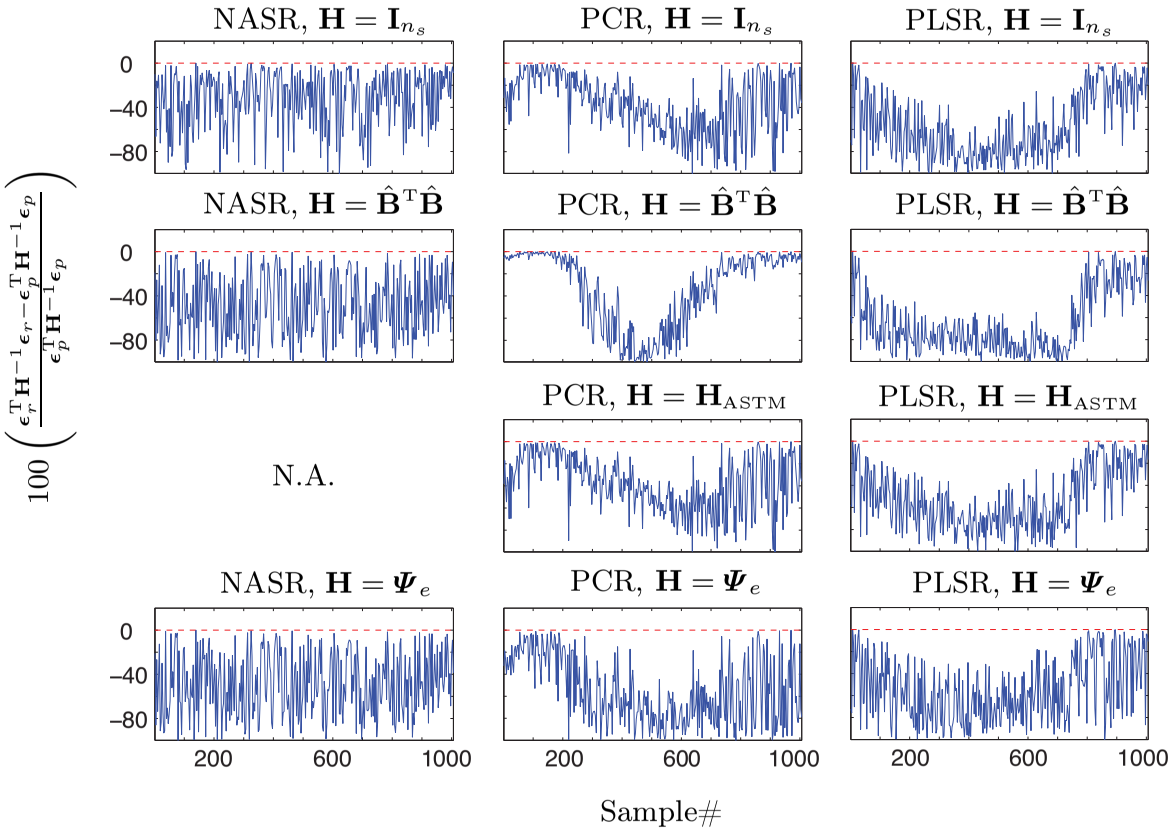


Fig. 5.3. Example 1: Proposition 5 is satisfied for NASR, PCR, and PLSR for all the cases of $\mathbf{H} = \mathbf{I}_{n_s}, \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}, \mathbf{H}_{\mathrm{ASTM}}$, and $\boldsymbol{\Psi}_e$. For NASR, the difference in weighted error norm is randomly distributed and iid with samples. In contrast to NASR, the difference in weighted error norm from PCR and PLSR is not randomly distributed and iid, because the bias in prediction error resulting from the error term $e1$ evolves gradually.

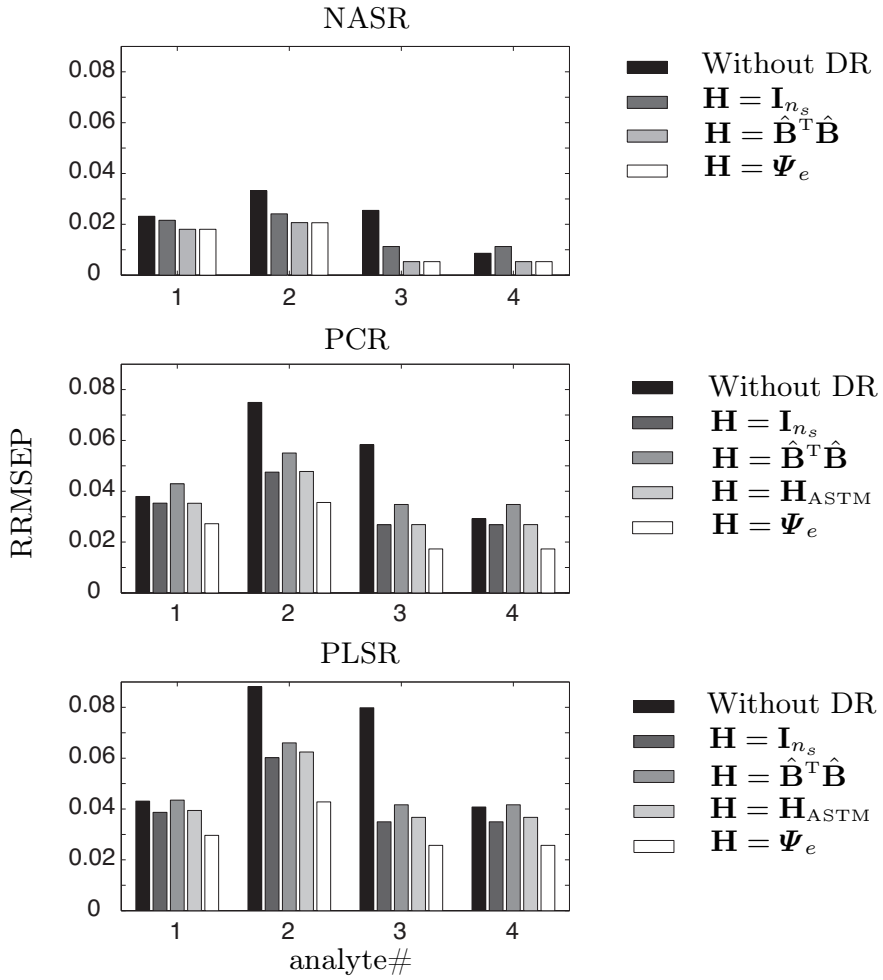


Fig. 5.4. Example 1: RRMSEP of the four analytes, averaged over 100 Monte Carlo simulations, for NASR, PCR, and PLSR, and different **H**.

5.3.2 Second example (experimental data): Use of closure equation

UV spectra of 114 samples of light gas oil and diesel fuel, used in the third example in Chapter 3, is used here with a different split for calibration and prediction. The properties of interest are the weight percentages of saturates, monoaromatics, diaromatics, and polyaromatics. The reference weight percentages are not actual measurements of saturates, monoaromatics, diaromatics, and polyaromatics, but rather values assigned based on a standard protocol, for example, monoaromatics are taken to be anything that elutes between a

specified range. Since the entire region is integrated, the weight percentages sum up to 100, leading to $n_g = 1$ closure equation:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} 100 \end{bmatrix} . \tag{5.17}$$

For illustration purposes, the data are split in such a way that the calibration data is representative of the prediction data. Ten samples from the first plant (out of 59), five from the second (out of 25), and five from the third (out of 30) are used for calibration, and the rest for prediction. This leads to a $[20 \times 572]$ matrix of calibration spectra $\tilde{\mathbf{X}}_c$, and a $[94 \times 572]$ matrix of prediction spectra $\tilde{\mathbf{X}}_p$. For each property of interest, four factors are chosen in its PCR and PLSR models. Since the pure-component spectra are unknown, NASR is not studied. Furthermore, the computation of $\boldsymbol{\Psi}_e$ requires the knowledge of $\boldsymbol{\epsilon}_p$. Since experimental reference measurements are always corrupted by measurement noise, the case of PCR or PLSR with $\mathbf{H} = \boldsymbol{\Psi}_e$ is also not included.

Discussion: The RRMSEP^{app} of the four response variables, averaged over 100 Monte Carlo simulations with different selection of data for calibration and validation, are shown in Fig. 5.5. In this example, $\mathbf{H} = \mathbf{I}_{n_s}$ leads to a small reduction in RRMSEP^{app} of the first response variable at the cost of a large increase in RRMSEP^{app} of the other three. Despite satisfying Proposition 5, if the predictions of all analytes were of equal importance in this example, DR with orthogonal projection may be judged to be of little merit. This example illustrates the importance of finding good estimates of the dispersion matrix of the prediction error. In contrast to Example 1, here RRMSEP^{app} with $\mathbf{H} = \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$ and $\mathbf{H} = \mathbf{H}_{\mathrm{ASTM}}$ are significantly better than those obtained with $\mathbf{H} = \mathbf{I}_{n_s}$.

5.4 Conclusions

DR applications commonly use $\mathbf{H} = \mathbf{I}_{n_s}$, i.e. all the variables involved in the balance equations are assumed to be predicted with equal uncertainty. However,

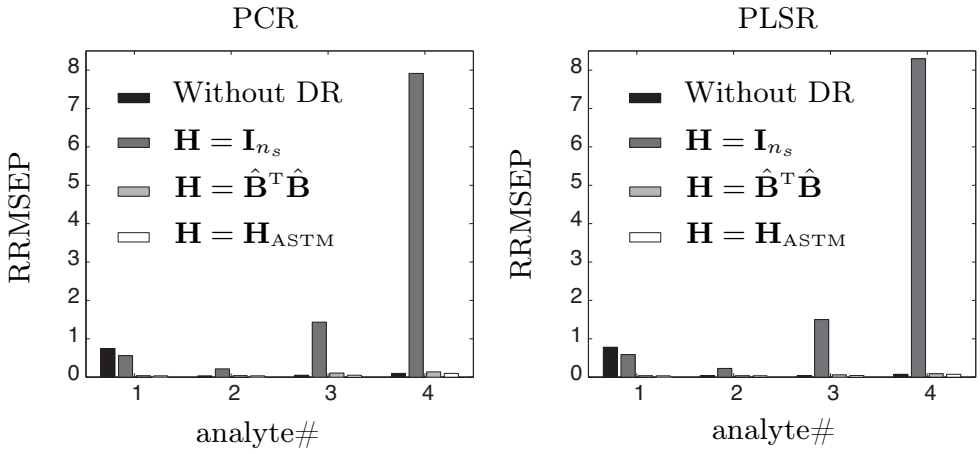


Fig. 5.5. Example 2: RRMSEP^{app} of the four analytes, averaged over 100 Monte Carlo simulations, for PCR and PLSR, and different $\mathbf{H}$.

if variables have widely differing prediction errors, DR may distribute the large prediction error amongst all variables, improving the prediction of the variables with large error at the expense of the variables with small errors. However, for $\mathbf{H} = \boldsymbol{\Psi}_e$, it is shown analytically that DR leads to reduction in RRMSEP for each variable.

If the bias of prediction errors is negligible compared to the variance, $\boldsymbol{\Psi}$ can be readily obtained as $\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{B}}$. If the off-diagonal elements of $\boldsymbol{\Psi}$ are negligible compared to the diagonal elements, $\boldsymbol{\Psi}$ can be approximated using analytical expressions of prediction confidence intervals. In PCR/PLSR, the bias in prediction is caused by subspace modeling error and drift. Furthermore, the off-diagonal elements of $\boldsymbol{\Psi}$ are non-zero since the same measurement noise propagates to the prediction errors for different analytes. In the author's opinion, estimation of $\boldsymbol{\Psi}$ in the presence of significant bias or off-diagonal elements must rely on "several" reference measurements. However, this may be expensive owing to the high costs of reference measurements. Alternately, drift subspace correction and latent subspace correction should be applied prior to data reconciliation.

6

Conclusions

6.1 Contributions

This section presents the main conclusions and contributions of the dissertation related to drift subspace correction using master/slave data, latent subspace correction using unlabeled data, and DR based on prior knowledge of linear dependencies.

Drift subspace correction:

- A framework for explicit drift-correction methods has been provided that consists of two main steps: (i) estimation of the drift subspace based upon different types of master/slave data, and (ii) correction of the calibration model for the estimated drift subspace by shrinkage or OP. The master/slave data are characterized as Type 1 if the linear operator is deduced from the knowledge of the master spectra, and as Type 2 if it is computed from the known master concentrations. The linear operator is constrained to satisfy equality of master and slave concentrations, since the drift affects the spectroscopic measurements but not the concentrations. This framework facilitates the evaluation and comparison of explicit drift-correction methods.
- The framework has been illustrated with different experimental data sets, correcting for effects of variations in temperature, moisture, particle size distribution, raw material, instrumental drift, and calibration transfer.

- It has been shown analytically that OP with full orthogonalization is equivalent to shinkage with a large meta-parameter, i.e. the drift subspace is shrunk completely. Under noise-free conditions, ICM and ECM with OP have been shown to be equivalent.

Latent subspace correction:

- Simulated and experimental data sets have been used to point out the usefulness and pitfalls of using unlabeled data. For drift-free measurements, the use of unlabeled data in addition to labeled data helps stabilize the latent subspaces in the calibration step, typically leading to a lower subspace modeling error. Unlabeled data can replace labeled data to some extent, thereby leading to an economic benefit. The advantage gained from using unlabeled data can be significant when the spectroscopic measurements have low signal-to-noise ratios. However, in the presence of drift, the use of unlabeled data can result in an increase in prediction error compared to that obtained with a model based on labeled data alone.
- The equivalence of different methods using unlabeled data in PCR and PLSR has been discussed. PCA-PLSR is shown to be equivalent to SO-PLSR. Hence, the only difference between NSO-PLSR and PCA-PLSR (or SO-PLSR) lies in the non-sequential estimation of factors in the NSO-PLSR and the sequential estimation in PCA-PLSR. For the examples considered, the difference in the RRMSEP with PCA-PLSR and NSO-PLSR is less than 0.05%. It is shown analytically that, with $r_{\mathrm{PLSR}} = r_{\mathrm{PCR}}$, all methods IT-PCR, PCA-PLSR, NSO-PLSR and SO-PLSR lead to the same regression vector. PCA-PLSR may be preferred over OF-based methods due to its simplicity.

Data reconciliation:

- It has been shown that a weighted l_2-norm of the prediction error vector is reduced for any estimate of the dispersion matrix of the prediction error. Moreover, RRMSEP for *each* analyte is reduced for "good" estimates of the dispersion matrix.
- If the bias components of prediction errors are negligible compared to the variance components, the dispersion matrix can be obtained directly from

the estimated regression vectors. If the off-diagonal elements of the dispersion matrix are negligible compared to the diagonal elements, the dispersion matrix can be obtained using analytical expressions of prediction confidence intervals. However, if the assumption does not hold and the dispersion matrix is poorly estimated, DR may improve the prediction of some analytes at the expense of others.

6.2 Perspectives

Soft sensors are indispensable tools for the process industries. These sensors facilitate process understanding and allow monitoring process operations to detect abnormal situations, thereby improving process reliability and leading to less wastage. However, the use of process analyzers in closed-loop control is still limited owing to their lack of robustness in harsh industrial environments. The proposed methods for drift subspace correction, latent subspace correction, and data reconciliation can contribute to increased robustness and reliability of predictions with few or no additional reference measurements. However, realizing the gains from the proposed methods is a matter of choosing the meta-parameters correctly. Ideally, this choice reflects the trade-offs between the different sources of prediction error. Since direct measurement of the error terms is in general not possible, finding suitable proxy estimates is an important subject for further investigation. Until such estimates are available, the choice of meta-parameters can only be guided by rules of thumb or heuristic criteria based on expert opinion, impeding the automated maintenance of models that is crucial for almost all advanced measurement systems in research and industry. Moreover, for soft sensors to be useful in QbD applications, better interpretability of the calibration models and mechanistic understanding of drift evolution will be indispensable.

The proposed methods can be extended to data-driven optimization schemes [86] for the purpose of process improvement. For example, in the context of the operational optimization of batch processes (e.g. by adjusting the trajectories of feed rates and/or process temperature [87, 88]) with terminal constraints (e.g. concentration of a side product less than a specified threshold at batch end), drift relates to batch-to-batch variations, which necessitates finding new opti-

mal input trajectories. The following challenges are anticipated. Firstly, based on historical batch data, the batch process should be approximated as a static linear (or nonlinear) calibration model between the parameterized input trajectories and the key variables at batch end. Secondly, based on on-line or at-line measurements (akin to master/slave data), the effect of batch variations on the calibration model should be evaluated, and a correction (akin to drift correction) applied to the calibration model to reduce sensitivity to the batch-to-batch variations. The proposed methods can also be extended to LV model predictive control [89], where the controlled errors are a function of LVs representing linear combinations of measurements, process variables, setpoints, and manipulated inputs. Subspace correction based on additional sources of information (e.g. unlabeled batch data or master/slave batch data) can lead to improved estimates of the LVs, and thus to a more reliable control action based on the corrected LVs.

A

Appendix A

The Appendix has space-inclusion conditions for noise-free spectra.

A.1 Space-inclusion conditions

Consider the following assumptions:

A1: $\mathbf{X}_c = \mathbf{Z}_c \mathbf{S}^{\mathrm{T}}$, $\mathbf{x}_p = \mathbf{S}\mathbf{z}_p$ (the calibration data and measured spectrum have no noise),

A2: $\mathrm{rank}(\mathbf{S}) = n_s$, $\mathrm{rank}\,(\mathbf{Z}_c) = q \leq n_s$ (pure-component spectra are of full rank but the concentration matrix can be rank deficient).

Proposition 7 *Let Assumptions A1–A2 hold. Then, the concentrations of the n_s analytes are predicted correctly from $\mathbf{x}_p$ using the inverse calibration model*

$$\hat{\mathbf{z}}_p^{\mathrm{T}} = \mathbf{x}_p^{\mathrm{T}} \hat{\mathbf{B}}, \quad \textit{where} \quad \hat{\mathbf{B}} = \mathbf{X}_c^{+} \mathbf{Z}_c \tag{A.1}$$

iff $\mathbf{x}_p \in \mathcal{R}\,(\mathbf{X}_c)$.

Proof of Proposition 7:

1) $\mathbf{x}_p \in \mathcal{R}\,(\mathbf{X}_c) \rightarrow$ correct prediction:

Let $\mathbf{X}_c$ be factorized using PCA and $\mathbf{Z}_c$ be factorized using the same scores:

$$\begin{aligned} \mathbf{X}_c &= \mathbf{Z}_c\,\mathbf{S}^{\mathrm{T}} = \mathbf{T}\,\mathbf{P}^{\mathrm{T}}, \\ \mathbf{Z}_c &= \mathbf{T}\,\mathbf{Q}^{\mathrm{T}} \end{aligned} \tag{A.2}$$

where $\mathbf{T}$ is the $[n_c \times q]$ scores matrix, and $\mathbf{P}$ and $\mathbf{Q}$ the loading matrices of size $[n_x \times q]$ and $[n_s \times q]$, respectively. From Eq. (A.2), $\mathbf{T}\,\mathbf{P}^{\mathrm{T}} = \mathbf{T}\,\mathbf{Q}^{\mathrm{T}}\,\mathbf{S}^{\mathrm{T}}$, which gives:

$$\mathbf{P}^{\mathrm{T}} = \mathbf{Q}^{\mathrm{T}}\,\mathbf{S}^{\mathrm{T}} \tag{A.3}$$

and

$$\hat{\mathbf{B}} = \mathbf{X}_c^{+}\,\mathbf{Z}_c = (\mathbf{P}^{\mathrm{T}+}\mathbf{T}^{+})\,(\mathbf{T}\,\mathbf{Q}^{\mathrm{T}}) = \mathbf{P}\,\mathbf{Q}^{\mathrm{T}}. \tag{A.4}$$

From space inclusion, $\mathbf{x}_p \in \mathcal{R}(\mathbf{X}_c)$, i.e. $\mathbf{x}_p^{\mathrm{T}} = \mathbf{g}_p^{\mathrm{T}}\,\mathbf{X}_c = \mathbf{g}_p^{\mathrm{T}}\,\mathbf{Z}_c\,\mathbf{S}^{\mathrm{T}}$. Since $\mathrm{rank}\,(\mathbf{S}) = n_s$ and $\mathbf{x}_p^{\mathrm{T}} = \mathbf{z}_p^{\mathrm{T}}\,\mathbf{S}^{\mathrm{T}}$, it follows that $\mathbf{z}_p^{\mathrm{T}} = \mathbf{g}_p^{\mathrm{T}}\,\mathbf{Z}_c$, which gives:

$$\hat{\mathbf{z}}_p^{\mathrm{T}} = \mathbf{x}_p^{\mathrm{T}}\,\hat{\mathbf{B}} = (\mathbf{g}_p^{\mathrm{T}}\,\mathbf{T}\,\mathbf{P}^{\mathrm{T}})\,(\mathbf{P}\,\mathbf{Q}^{\mathrm{T}}) = \mathbf{g}_p^{\mathrm{T}}\,\mathbf{T}\,\mathbf{Q}^{\mathrm{T}} = \mathbf{g}_p^{\mathrm{T}}\,\mathbf{Z}_c = \mathbf{z}_p^{\mathrm{T}}. \tag{A.5}$$

2) Correct prediction $\rightarrow \mathbf{x}_p \in \mathcal{R}(\mathbf{X}_c)$*:*

Using $\mathbf{x}_p^{\mathrm{T}} = \mathbf{z}_p^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}$, $\hat{\mathbf{z}}_p^{\mathrm{T}} = \mathbf{z}_p^{\mathrm{T}}$ and Eqs. (A.3) and (A.4),

$$\mathbf{x}_p^{\mathrm{T}} = \hat{\mathbf{z}}_p^{\mathrm{T}}\,\mathbf{S}^{\mathrm{T}} = \mathbf{x}_p^{\mathrm{T}}\,\mathbf{P}\,\mathbf{Q}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}} = \mathbf{x}_p^{\mathrm{T}}\,\mathbf{P}\,\mathbf{P}^{\mathrm{T}}, \tag{A.6}$$

i.e. $\mathbf{x}_p^{\mathrm{T}} = \mathbf{x}_p^{\mathrm{T}}\,\mathbf{X}_c^{+}\,\mathbf{X}_c$, implying space inclusion. □

Proposition 8 *Let Assumptions A1–A2 hold. Then, the concentrations of the* n_k *analytes are predicted correctly from* $\mathbf{x}_p$ *using the inverse calibration model*

$$\hat{\mathbf{z}}_{p,k}^{\mathrm{T}} = \mathbf{x}_p^{\mathrm{T}}\,\hat{\mathbf{B}}_k, \quad \textit{where} \quad \hat{\mathbf{B}}_k = \mathbf{X}_c^{+}\,\mathbf{Z}_{c,k} \tag{A.7}$$

if $\mathbf{x}_p \in \mathcal{R}(\mathbf{X}_c)$.

Proof of Proposition 8:

1) $\mathbf{x}_p \in \mathcal{R}(\mathbf{X}_c) \rightarrow$ *correct prediction:*

Let $\mathbf{Z}_c = [\mathbf{Z}_{c,k}\,\mathbf{Z}_{c,u}]$ and $\hat{\mathbf{B}} = \mathbf{X}_c^{+}\,\mathbf{Z}_c = [\hat{\mathbf{B}}_k\,\hat{\mathbf{B}}_u]$. $\mathbf{Z}_{c,k} = \mathbf{Z}_c\,\mathbf{J}_k$, $\hat{\mathbf{B}}_k = \hat{\mathbf{B}}\,\mathbf{J}_k$, $\mathbf{Z}_{c,u} = \mathbf{Z}_c\,\mathbf{J}_u$ and $\hat{\mathbf{B}}_u = \hat{\mathbf{B}}\,\mathbf{J}_u$, where $\mathbf{J}_k \equiv \begin{bmatrix} \mathbf{I}_{n_k} \\ \mathbf{0}_{n_u \times n_k} \end{bmatrix}$ and $\mathbf{J}_u \equiv \begin{bmatrix} \mathbf{0}_{n_k \times n_u} \\ \mathbf{I}_{n_u} \end{bmatrix}$. Following the steps leading to Eq. (A.5),

$$\hat{\mathbf{z}}_p^{\mathrm{T}}\mathbf{J}_k = \mathbf{x}_p^{\mathrm{T}}\,\hat{\mathbf{B}}\mathbf{J}_k = \mathbf{z}_p^{\mathrm{T}}\mathbf{J}_k = \mathbf{z}_{p,k}^{\mathrm{T}}. \tag{A.8}$$

Hence, $\hat{\mathbf{z}}_{p,k}^{\mathrm{T}} = \mathbf{z}_{p,k}^{\mathrm{T}}$.

2) Correct prediction $\nrightarrow$ $\mathbf{x}_p \in \mathcal{R}(\mathbf{X}_c)$:

One example will suffice to illustrate that correct prediction does not imply space inclusion. Choose,

$$\mathbf{z}_p^{\mathrm{T}} = \mathbf{g}_p^{\mathrm{T}}\,\mathbf{Z}_c + \mathbf{z}_p^{\perp\mathrm{T}}, \quad \mathbf{z}_p^{\perp\mathrm{T}} = [\mathbf{0}_{n_k}^{\mathrm{T}} \;\; \mathbf{m}^{\mathrm{T}}]\,, \tag{A.9}$$

where $\mathbf{z}_p^{\perp}$ is the orthogonal complement to $\mathcal{R}(\mathbf{Z}_c)$, and $\mathbf{m} \neq \mathbf{0}_{n_u}$ is an n_u-dimensional vector. Analyte concentrations are predicted as,

$$\begin{aligned} \hat{\mathbf{z}}_{p,k}^{\mathrm{T}} = &\quad \mathbf{x}_p^{\mathrm{T}}\,\hat{\mathbf{B}}_k = \mathbf{z}_p^{\mathrm{T}}\,\mathbf{S}^{\mathrm{T}}\,\hat{\mathbf{B}}_k = (\mathbf{g}_p^{\mathrm{T}}\,\mathbf{Z}_c + \mathbf{z}_p^{\perp\mathrm{T}})\,\mathbf{S}^{\mathrm{T}}\,\mathbf{P}\mathbf{Q}^{\mathrm{T}}\mathbf{J}_k \\ = &\quad \mathbf{g}_p^{\mathrm{T}}\,\mathbf{T}\mathbf{Q}^{\mathrm{T}}\,\mathbf{S}^{\mathrm{T}}\,\mathbf{P}\mathbf{Q}^{\mathrm{T}}\mathbf{J}_k + \mathbf{z}_p^{\perp\mathrm{T}}\,\mathbf{S}^{\mathrm{T}}\,\mathbf{P}\mathbf{Q}^{\mathrm{T}}\mathbf{J}_k\,. \end{aligned} \tag{A.10}$$

Using $\mathbf{Q}^{\mathrm{T}}\,\mathbf{S}^{\mathrm{T}}\,\mathbf{P}\,\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{\mathrm{T}}$ (see Eq. (A.3)) and $\mathbf{z}_p^{\perp\mathrm{T}}\,\mathbf{S}^{\mathrm{T}} = \mathbf{m}^{\mathrm{T}}\,\mathbf{S}_u^{\mathrm{T}}$,

$$\begin{aligned} \hat{\mathbf{z}}_{p,k}^{\mathrm{T}} = &\quad \mathbf{g}_p^{\mathrm{T}}\,\mathbf{T}\mathbf{Q}^{\mathrm{T}}\mathbf{J}_k + \mathbf{m}^{\mathrm{T}}\,\mathbf{S}_u^{\mathrm{T}}\mathbf{P}\mathbf{Q}^{\mathrm{T}}\mathbf{J}_k = \mathbf{g}_p^{\mathrm{T}}\,\mathbf{Z}_c\mathbf{J}_k + \mathbf{m}^{\mathrm{T}}\,\mathbf{S}_u^{\mathrm{T}}\hat{\mathbf{B}}_k \\ = &\quad \mathbf{z}_{p,k}^{\mathrm{T}} + \mathbf{m}^{\mathrm{T}}\,\mathbf{S}_u^{\mathrm{T}}\hat{\mathbf{B}}_k\,. \end{aligned} \tag{A.11}$$

It is easy to construct a synthetic example with $\mathbf{m}$, $\mathbf{Z}_c$, and $\mathbf{S}$, that leads to correct predictions, i.e. $\mathbf{m}^{\mathrm{T}}\,\mathbf{S}_u^{\mathrm{T}}\hat{\mathbf{B}}_k = \mathbf{0}_{n_k}^{\mathrm{T}}$. The corresponding spectra is given as,

$$\mathbf{x}_p^{\mathrm{T}} = (\mathbf{g}_p^{\mathrm{T}}\,\mathbf{Z}_c + \mathbf{z}_p^{\perp\mathrm{T}})\,\mathbf{S}^{\mathrm{T}} = \mathbf{g}_p^{\mathrm{T}}\,\mathbf{Z}_c\,\mathbf{S}^{\mathrm{T}} + \mathbf{z}_p^{\perp\mathrm{T}}\,\mathbf{S}^{\mathrm{T}}\,. \tag{A.12}$$

The first term $\mathbf{g}_p^{\mathrm{T}}\,\mathbf{X}_c \in \mathcal{R}(\mathbf{X}_c)$, but the second term $\mathbf{z}_p^{\perp\mathrm{T}}\,\mathbf{S}^{\mathrm{T}} \notin \mathcal{R}(\mathbf{X}_c)$. □

B

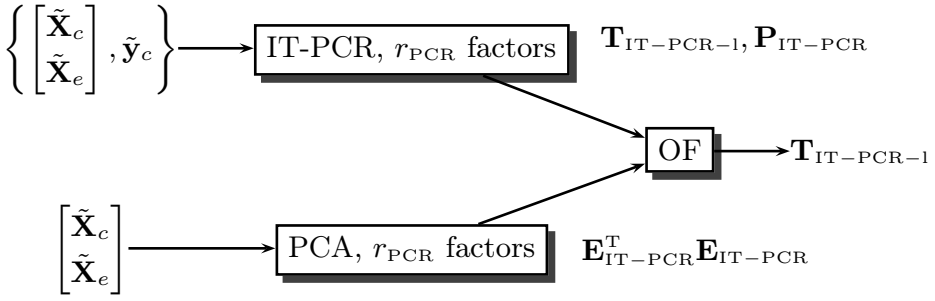
Appendix B

The Appendix provides discussions on OF-based PCR and PLSR.

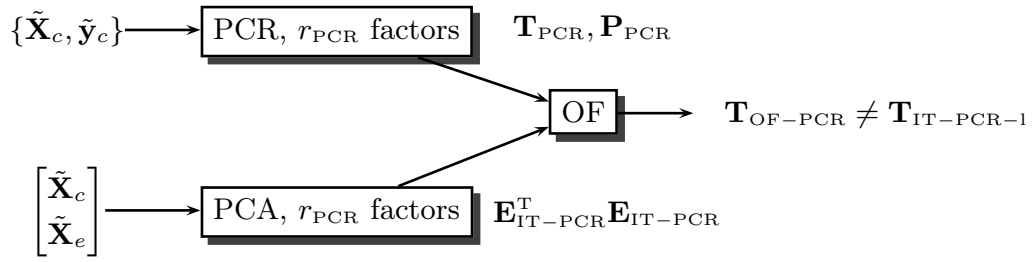
B.1 Discussion on OF-based PCR

Ergon and Esbensen showed that the OF-based PCR is the same as IT-PCR [67]. However, this claim needs to be clarified as the following two schematics can be considered for OF-based PCR:

Schematic 1:



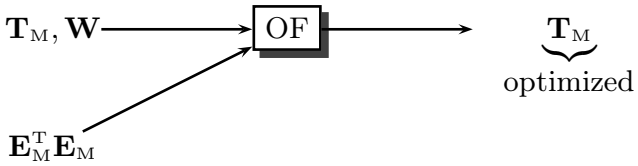
Schematic 2:



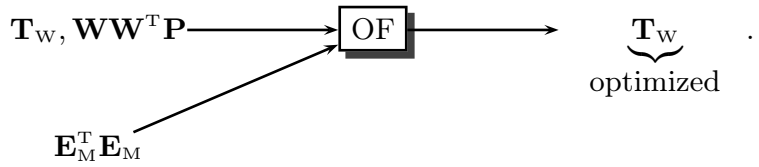
Schematic 1 was considered in [67]. Here, OF does not alter $\mathbf{T}_{\mathrm{IT-PCR-1}}$, implying that the scores in IT-PCR are already optimized with respect to the loadings and error covariance. However, strictly speaking, Schematic 1 represents an OF-based IT-PCR rather than an OF-based PCR. In Schematic 2, the output of OF is $\mathbf{T}_{\mathrm{OF-PCR}}$ with, in general, $\mathbf{T}_{\mathrm{OF-PCR}} \neq \mathbf{T}_{\mathrm{IT-PCR-1}}$. Since Schematic 1 uses stabilized estimates of the scores and loadings, it is preferred over Schematic 2.

B.2 Discussion on OF-based PLSR

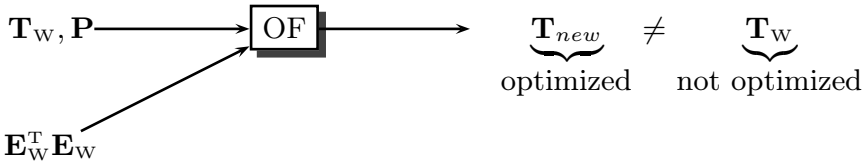
Recently, several authors have debated whether Wold's PLSR is inconsistent with respect to the model space used in calibration and prediction, while Martens' PLSR is consistent [18, 90–92]. In the context of OF, this inconsistency can be noted in the fact that the scores in Martens' PLSR and Ergon's PLSR are optimized with respect to its loadings and error covariance, while scores in Wold's PLSR are not. This is shown in the following schematic for Martens' PLSR,



and for Ergon's PLSR,



Since $\mathbf{T}_{\mathrm{M}}$ is already optimized with respect to $\mathbf{W}$ and $\mathbf{E}_{\mathrm{M}}$, OF does not alter $\mathbf{T}_{\mathrm{M}}$. Similarly, since $\mathbf{T}_{\mathrm{W}}$ is already optimized with respect to $\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{P}$ and $\mathbf{E}_{\mathrm{M}}$, OF does not alter $\mathbf{T}_{\mathrm{W}}$. However, in the case of Wold's PLSR, $\mathbf{T}_{\mathrm{W}}$ is not optimized with respect to $\mathbf{P}$ and $\mathbf{E}_{\mathrm{W}}$ as shown in the following schematic:



The different PLSR models with unlabeled data presented in Chapter 4 are valid only for consistent PLSR models, i.e. Martens' PLSR or Ergon's PLSR.

References

[1] FDA. Process analytical technology (PAT) initiative. Available from: http://www.fda.gov/cder/OPS/PAT.htm. 2005.

[2] Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*. 2009;33(4):795–814.

[3] Blanco M, Alcalá M. Use of near-infrared spectroscopy for off-line measurements in the pharmaceutical industry. In: *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries*, edited by Bakeev KA, vol. 1, pp. 362–391. Oxford, UK: Blackwell Publishing, 1st ed. 2005.

[4] Belsey DA, Kuh E, Welsch R. *Regression Diagnostics*. New York, USA: John Wiley & Sons. 1980.

[5] Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55–67.

[6] Cleveland WS, Devlin SJ. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*. 1988;83(403):596–610.

[7] Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267–288.

[8] Friedman JH, Stuetzle W. Projection pursuit regression. *Journal of the American Statistical Association*. 1981;76(376):817–823.

[9] Friedman JH. Multivariate adaptive regression splines. *Annals of Statistics*. 1991;19(1):1–67.

[10] Haykin S. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice Hall PTR. 1998.

[11] Thissen U, Pepers M, Üstün B, Melssen WJ, Buydens LMC. Comparing support vector machines to PLS for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems*. 2004;73(2):169–179.

[12] Chen T, Morris J, Martin E. Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems*. 2007;87(1):59–71.

[13] Marbach R. A new method for multivariate calibration. *Journal of Near Infrared Spectroscopy*. 2005;13(5):241–254.

[14] Wold S, Sjöström M, Eriksson L. PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*. 2001; 58(2):109–130.

[15] Næs T, Isaksson T, Fearn T, Davies T. *A User-Friendly Guide to Multivariate Calibration and Classification*. Chichester, UK: NIR Publications. 2002.

[16] Martens H. Reliable and relevant modelling of real world data: A personal account of the development of PLS Regression. *Chemometrics and Intelligent Laboratory Systems*. 2001;58(2):85–95.

[17] Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*. 1984;5(3):735–743.

[18] Ergon R. Re-interpretation of NIPALS results solves PLSR inconsistency problem. *Journal of Chemometrics*. 2008;23(2):72–75.

[19] Helland IS. On the structure of partial least squares regression. *Communications in Statistics - Simulation and Computation*. 1988;17(2):581–607.

[20] Martens H, Næs T. *Multivariate Calibration*. Chichester, UK: John Wiley & Sons. 1989.

[21] Faber NM. The X-residuals calculated by partial least squares are problematic for uncertainty estimation. *Chemometrics and Intelligent Laboratory Systems*. 2009;96(2):264–265.

[22] Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*. 1983; 78(382):316–331.

[23] Anderssen E, Dyrstad K, Westad F, Martens H. Reducing over-optimism in variable selection by cross-model validation. *Chemometrics and Intelligent Laboratory Systems*. 2006;84(1-2):69–74.

[24] Lorber A, Faber K, Kowalski BR. Net analyte signal calculation in multivariate calibration. *Analytical Chemistry*. 1997;69(8):1620–1626.

[25] Brown CD. Discordance between net analyte signal theory and practical multivariate calibration. *Analytical Chemistry*. 2004;76(15):4364–4373.

[26] Estienne F, Despagne F, Walczak B, de Noord OE, Massart DL. A comparison of multivariate calibration techniques applied to experimental NIR data sets: Part III: Robustness against instrumental perturbation conditions. *Chemometrics and Intelligent Laboratory Systems*. 2004;73(2):207–218.

[27] Larrechi MS, Callao MP. Strategy for introducing NIR spectroscopy and multivariate calibration techniques in industry. *TrAC Trends in Analytical Chemistry*. 2003;22(9):634–640.

[28] Saiz-Abajo MJ, Mevik BH, Segtnan VH, Næs T. Ensemble methods and data augmentation by noise addition applied to the analysis of spectroscopic data. *Analytica Chimica Acta*. 2005;533(2):147–159.

[29] Wise B, Gallagher N, Butler S, White D, Barna G. Development and benchmarking of multivariate statistical process control tools for a semiconductor etch process: Improving robustness through model updating. In: *IFAC ADCHEM'97*. Banff, Canada. 1997; pp. 78–83.

[30] Rocco D. Examination of some misconceptions about near-infrared analysis. *Applied Spectroscopy*. 1995;49:67–75.

[31] Artursson T, Eklöv T, Lundström I, Mårtensson P, Sjöström M, Holmberg M. Drift correction for gas sensors using multivariate methods. *Journal of Chemometrics*. 2000;14(5-6):711–723.

[32] Hansen PW. Pre-processing method minimizing the need for reference analyses. *Journal of Chemometrics*. 2001;15(2):123–131.

[33] Martens H, Høy M, Wise MB, Bro R, Brockhoff BP. Pre-whitening of data by covariance-weighted pre-processing. *Journal of Chemometrics*. 2003; 17(3):153–165.

[34] Wise BM, Gallagher NB, Bro R, Shaver JM, Windig W, Koch RS. PLS_Toolbox for use with MATLAB. 2008.

[35] Roger JM, Chauchard F, Bellon-Maurel V. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. *Chemometrics and Intelligent Laboratory Systems*. 2003;66(2):191–204.

[36] Andrew A, Fearn T. Transfer by orthogonal projection: Making near-infrared calibrations robust to between-instrument variation. *Chemometrics and Intelligent Laboratory Systems*. 2004;72(1):51–56.

[37] Zeaiter M, Roger JM, Bellon-Maurel V. Dynamic orthogonal projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR-based monitoring of wine fermentations. *Chemometrics and Intelligent Laboratory Systems*. 2006;80(2):227–235.

[38] Dabros M, Amrhein M, Gujral P, von Stockar U. On-line recalibration of spectral measurements using metabolite injections and dynamic orthogonal projection. *Applied Spectroscopy*. 2007;61(5):507–513.

[39] Segtnan VH, Mevik BH, Isaksson T, Næs T. Low-cost approaches to robust temperature compensation in near-infrared calibration and prediction situations. *Applied Spectroscopy*. 2005;59:816–825.

[40] Zhu Y, Fearn T, Samuel D, Dhar A, Hameed O, Bown SG, Lovat LB. Error removal by orthogonal subtraction (EROS): A customised pre-treatment for spectroscopic data. *Journal of Chemometrics*. 2008;22(2):130–134.

[41] Jianguo S. Statistical analysis of NIR data: Data pretreatment. *Journal of Chemometrics*. 1997;11(6):525–532.

[42] Pierna JAF, Chauchard F, Preys S, Roger JM, Galtier O, Baeten V, Dardenne P. How to build a robust model against perturbation factors with only a few reference values: A chemometric challenge at 'Chimiométrie 2007'. *Chemometrics and Intelligent Laboratory Systems*. 2010;doi:10.1016/j.chemolab.2010.05.015.

[43] Faber K, Kowalski BR. Modification of Malinowski's F-test for abstract factor analysis applied to the Quail Roost II data sets. *Journal of Chemometrics*. 1997;11(1):53–72.

[44] Malinowski ER. Statistical F-tests for abstract factor analysis and target testing. *Journal of Chemometrics*. 1989;3(1):49–60.

[45] Kritchman S, Nadler B. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*. 2008;94(1):19–32.

[46] Krzanowski WJ. Between-group comparison of principal components - some sampling results. *Journal of Statistical Computation and Simulation*. 1982;15(2):141–154.

[47] Wulfert F, Kok WT, Smilde AK. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Analytical Chemistry*. 1998;70(9):1761–1767.

[48] Wentzell PD, Andrews DT, Walsh JM, Cooley JM, Spencer P. Estimation of hydrocarbon types in light gas oils and diesel fuels by ultraviolet absorption spectroscopy and multivariate calibration (the data are available at http://myweb.dal.ca/pdwentze/downloads.html). *Canadian Journal of Chemistry*. 1999;77(3):391–400.

[49] The data are available at the Eigenvector website: http://www.eigenvector.com/Data/Corn/.

[50] Thijssen PC, Wolfrum SM, Kateman G, Smit HC. A Kalman filter for calibration, evaluation of unknown samples and quality control in drifting

systems: Part 1. Theory and Simulations. *Analytica Chimica Acta*. 1984; 156:87–101.

[51] Brown SD. The Kalman filter in analytical chemistry. *Analytica Chimica Acta*. 1986;181:1–26.

[52] Junde L, Ke Y, Furong G. Process similarity and developing new process models through migration. *AIChE Journal*. 2009;55(9):2318–2328.

[53] Haaland DM, Melgaard DK. New prediction-augmented classical least-squares (PACLS) methods: Application to unmodeled interferents. *Applied Spectroscopy*. 2000;54(9):1303–1312.

[54] Haaland DM, Melgaard DK. New classical least-squares/partial least-squares hybrid algorithm for spectral analyses. *Applied Spectroscopy*. 2001; 55(1):1–8.

[55] Wold S, Antti H, Lindgren F, Ohman J. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*. 1998;44(1-2):175–185.

[56] Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*. 2002;16(3):119–128.

[57] Svensson O, Kourti T, MacGregor JF. An investigation of orthogonal signal correction algorithms and their characteristics. *Journal of Chemometrics*. 2002;16(4):176–188.

[58] Vogt F, Steiner H, Booksh K, Mizaikoff B. Chemometric correction of drift effects in optical spectra. *Applied Spectroscopy*. 2004;58(6):683–692.

[59] Helland IS, Næs T, Isaksson T. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*. 1995;29(2):233–241.

[60] Liang F, Mukherjee S, West M. The use of unlabeled data in predictive modeling. *Statistical Science*. 2007;22(2):189–205.

[61] Zien A, Brefeld U, Scheffer T. Transductive support vector machines for structured variables. In: *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, edited by Ghahramani Z. ACM Press, New York, USA. 2007; pp. 1183–1190.

[62] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In: *Annual Conference on Computational Learning Theory*. ACM Press, New York, USA. 1998; pp. 92–100.

[63] Gaksoo L, Cheong Hee P. Semi-supervised dimension reduction using graph-based discriminant analysis. In: *Ninth IEEE International Conference on Computer and Information Technology*, vol. 2. Xiamen: IEEE Computer Society. 2009; pp. 9–13.

[64] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*. 1977;39(1):1–38.

[65] Isaksson T, Næs T. Selection of samples for calibration in near-infrared spectroscopy. Part II: Selection based on spectral measurements. *Applied Spectroscopy*. 1990;44:1152–1158.

[66] Thomas EV. Incorporating auxiliary predictor variation in principal component regression models. *Journal of Chemometrics*. 1995;9(6):471–481.

[67] Ergon R, Esbensen HK. PCR/PLSR optimization based on noise covariance estimation and Kalman filtering theory. *Journal of Chemometrics*. 2002;16(8-10):401–407.

[68] Ergon R. Constrained numerical optimization of PCR/PLSR predictors. *Chemometrics and Intelligent Laboratory Systems*. 2003;65(2):293–303.

[69] Bernstein DS. *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*. NJ, USA: Princeton University Press. 2005.

[70] Osten DW, Kowalski BR. Background detection and correction in multicomponent analysis. *Analytical Chemistry*. 1985;57(4):908–917.

[71] Wentzell PD, Andrews DT, Kowalski BR. Maximum likelihood multivariate calibration. *Analytical Chemistry*. 1997;69:2299–2311.

[72] Voet Hvd. Pseudo-degrees of freedom for complex predictive models: The example of partial least squares. *Journal of Chemometrics*. 1999;13:195–208.

[73] Wentzell PD, Montoto LV. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*. 2003; 65(2):257–279.

[74] The data are available at the IDRC website: http://www.idrc-chambersburg.org/shootout_2002.htm.

[75] Winning H, Larsen FH, Bro R, Engelsen SB. Quantitative analysis of NMR spectra with chemometrics. *Journal of Magnetic Resonance*. 2008; 190(1):26–32.

[76] Dabros M, Amrhein M, Bonvin D, Marison IW, Stockar Uv. Data reconciliation of concentration estimates from mid-infrared and dielectric spectral measurements for improved on-line monitoring of bioprocesses. *Biotechnology Progress*. 2009;25(2):578–588.

[77] Romagnoli JA, Sanchez MC. *Data Processing and Reconciliation for Chemical Process Operations*, vol. 2 of *Process systems engineering*. San Diego, USA: Academic Press. 2000.

[78] Wang NS, Stephanopoulos G. Application of macroscopic balances to the identification of gross measurement errors. *Biotechnology and Bioengineering*. 1983;25(9):2177–2208.

[79] Lange HC, Heijnen JJ. Statistical reconciliation of the elemental and molecular biomass composition of Saccharomyces cerevisiae. *Biotechnology and Bioengineering*. 2001;75(3):334–344.

[80] Carnicer M, Baumann K, Töplitz I, Sanchez-Ferrando F, Mattanovich D, Ferrer P, Albiol J. Macromolecular and elemental composition analysis and extracellular metabolite balances of Pichia pastoris growing at different oxygen levels. *Microbial Cell Factories*. 2009;8(1):65.

[81] Faber NM, Song XH, Hopke PK. Sample-specific standard error of prediction for partial least squares regression. *TrAC Trends in Analytical Chemistry*. 2003;22(5):330–334.

[82] Høy M, Steen K, Martens H. Review of partial least squares regression prediction error in Unscrambler. *Chemometrics and Intelligent Laboratory Systems*. 1998;44(1-2):123–133.

[83] Pierna JAF, Jin L, Wahl F, Faber NM, Massart DL. Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error. *Chemometrics and Intelligent Laboratory Systems*. 2003;65(2):281–291.

[84] American Society for Testing and Materials, Annual book of ASTM standards 03.06, E1655: Standard practices for infrared, multivariate, quantitative analysis, ASTM International, West Conshohocken, Pennsylvania, USA, 1998.

[85] Amrhein M, Srinivasan B, Bonvin D, Schumacher MM. On the rank deficiency and rank augmentation of the spectral measurement matrix. *Chemometrics and Intelligent Laboratory Systems*. 1996;33(1):17–33.

[86] Bonvin D, Srinivasan B, Ruppen D. Dynamic optimization in the batch chemical industry. In: *International Conference on Chemical Process Control (CPC VI), AIChE Symposium*, vol. 98 of *Series 326*. 2001; pp. 255–273.

[87] François G. Measurement-based run-to-run optimization of batch processes: Application to industrial acrylamide copolymerization. Ph.D. thesis, EPF Lausanne, Switzerland. 2004.

[88] Kourti T. Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*. 2005;19(4):213–246.

[89] Flores-Cerrillo J, MacGregor JF. Latent variable MPC for trajectory tracking in batch processes. *Journal of Process Control*. 2005;15(6):651–663.

[90] Pell RJ, Ramos LS, Manne R. The model space in partial least squares regression. *Journal of Chemometrics*. 2007;21(3-4):165–172.

[91] Bro R, Eldén L. PLS works. *Journal of Chemometrics*. 2009;23:69–71.

[92] Wold S, Høy M, Martens H, Trygg J, Westad F, MacGregor JF, Wise BM. The PLS model space revisited. *Journal of Chemometrics*. 2009;23:67–68.

Pamandeep Gujral

Date of birth: 02/04/1981, Nationality: Indian

EDUCATION

PhD	Computer & Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland	2005-2010
Bachelors & Masters	Elec. Engg. & Communication Systems, Indian Institute of Technology in Madras, India	1999-2004

EXPERIENCE

Engineer	Online Control, Lausanne, Switzerland	June-Aug 2010
Engineer	Texas Instruments, Bangalore, India	2004-2005

PUBLICATIONS

- Gujral P, Amrhein M, Wise BM, Bonvin D. On multivariate calibration with unlabeled data, submitted to the *Journal of Chemometrics*.
- Gujral P, Amrhein M, Wise BM, Bonvin D. Framework for explicit drift correction in multivariate calibration models, *Journal of Chemometrics*, 2010; 24:534–543.
- Gujral P, Wise BM, Amrhein M, Bonvin D. Partial least-squares regression with unlabeled data. In: *6th International Conference on Partial Least Squares and Related Methods*. Beijing, China, 2009; vol. 1, pp. 102–105.
- Gujral P, Amrhein A, Bonvin D, Vallée JP, Montet X, Michoux N. Classification of magnetic resonance images from rabbit renal perfusion, *Chemometrics and Intelligent Laboratory*, 2009; 98:173–181.
- Gujral P, Amrhein M, Bonvin D. Drift correction in multivariate calibration models using on-line reference measurements, *Analytica Chimica Acta*, 2009; 642:27–32.
- Gujral P, Wise BM, Amrhein M, Bonvin D. Principal component regression with unlabeled data. In: *11th Scandinavian Symposium on Chemometrics*. Loen, Norway, 2009; vol. 1, pp. 79.
- Gujral P, Amrhein M, Wise BM, Guzman E, Chivala D, Bonvin D. Correction of systematic disturbances in latent-variable calibration models. In: *11th Scandinavian Symposium on Chemometrics*. Loen, Norway, 2009; vol. 1, pp. 36.

- Gujral P, Amrhein M, Bonvin D. Measurement-based drift correction in spectroscopic calibration models. In: *11th Conference on Chemometrics in Analytical Chemistry*. Montpellier, France, 2008; vol. 1, pp. 117–121.
- Dabros M, Amrhein A, Gujral P, von Stockar U. On-line data reconciliation of mid-infrared and dielectric spectral measurements for the estimation of analyte and biomass concentrations in microbial fermentations. In: *Journal of Biotechnology*. Barcelona, Spain, 2007; vol. 131, pp. S174.
- Dabros M, Amrhein M, Gujral P, Marison I, von Stockar U. On-line recalibration of spectral measurements using metabolite injections and dynamic orthogonal projection, *Applied Spectroscopy*, 2007; 61:507–513.

Index