A CHEMOMETRICIAN'S GUIDE TO TRANSFER LEARNING

Preprint submitted to Journal of Chemometrics

Ramin Nikzad-Langerodi, Dr.rer.nat
Software Competence Center Hagenberg
Softwarepark 32a
Hagenberg, 4232, Austria
ramin.nikzad-langerodi@scch.com

Erik Andries, PhD
Central New Mexico Community College (CNM)
900 University Blvd SE
Albuquerque, NM, USA

Center for Advanced Research Computing
University of New Mexico
Albuquerque NM, USA erik.andries@gmail.com

October 12, 2021

ABSTRACT

Transfer learning (TL), the sub-discipline of machine learning devoted to learning from different domains, has gained increasing attention over the past decade. With the current contribution, we aim at giving a concise overview on theory, concepts and applications of TL from a chemometrician's perspective and draw some connections to previous work on calibration model updating/adaptation and calibration transfer. Furthermore, we provide a demonstration of the application of TL in analytical chemistry and discuss the benefits and challenges associate with its use for real-world problems. We conclude the paper by discussing some open problems and by contemplating on future research directions.

***K*eywords** transfer learning, domain adaptation, covariate shift, model maintenance, calibration transfer, data fusion

1 Introduction

According to definitions given by Pan *et al.*, transfer learning (TL) aims at leveraging knowledge gained when learning to solve one task to solve another, related task [1]. TL has gained increasing attention over the past decade, especially in the computer vision domain, as deep neural networks (DNN) trained on massive amounts of (image) data have become publicly available [2]. Such pre-trained models do not usually give satisfactory results when applied in new domains. For instance, a (DNN) model trained on images of objects (e.g. cars and bicycles) on a white background will run into trouble classifying these objects correctly on images from real-world scenes. Likewise, a chemometric model for the determination of the concentration of some analyte in water (from a spectroscopic signal) will typically give wrong results when applied to the determination of the same analyte e.g. in blood. As we will see later, TL provides mechanisms to adapt such models to this type of changes.

A large number of TL approaches have been developed over the past decades. Our intention is to provide the average reader, with a background in chemometrics, a concise overview on the theory and (in our view) the most important concepts that are relevant to applications in analytical chemistry rather than to give an exhaustive overview of the current state-of-the-art (SoA) in the rapidly evolving field of transfer learning. For the interested reader, we refer to the excellent reviews by Pan *et al.* and Weiss *el al.* [1] (for a general overview on TL) and Zhuang *et al.* [3] for an in-depth discussion of the current SoA [4]. Also, we will restrict our considerations mostly to application of TL to multivariate regression/calibration problems in spectroscopy.

We will continue by introducing some important notation and definitions from the TL field in section 2.1. We will then introduce the theory of learning from different domains in section 2.2. In section 2.3 we will explain some concepts and how these pertain to typical applications in analytical chemistry. In section 3 we will highlight previous work on TL in chemometrics and in section 4 we will showcase the use of TL on simulated and real-world data including some

discussion on the benefits, challenges and pitfalls of applying TL in practice. In section 5 we will provide some thoughts on where we see opportunities for the application of TL in chemometrics and how to address some open challenges in future work, and section 6 concludes the paper.

2 Theory

2.1 Notation

We will follow standard notation used in chemometrics, with upper and lower case boldface symbols (e.g. $\mathbf{X}$ and $\mathbf{x}$) denoting matrices and vectors, respectively. Unless otherwise stated, upper and lower case roman letters will be used to denote random variables (e.g. X) and scalars (e.g. x), and vectors are column vectors. By $^{\mathrm{T}}$ and $^{-1}$ we denote the transpose and inverse operation, respectively. $\mathbf{I}$ and $\mathbf{1}$ will be used to indicate an identity matrix and a column vector of ones, respectively, of appropriate size. With $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ we denote the ℓ_2- and the Frobenius norm (the Frobenius norm is an extension of the standard ℓ_2 norm except that an $n \times p$ matrix is treated as a vector of length np). The superscripted † in $\mathbf{X}^\dagger$ denotes the Moore-Penrose inverse of the matrix $\mathbf{X}$. Comma and semicolon notation are used to denote horizontal and vertical concatenation or stacking of matrices and/or scalars, e.g. $[\mathbf{x}, \mathbf{y}]$ and $[\mathbf{x}; \mathbf{y}]$.

We further follow the definitions given in [5] and let a *domain* consist of a marginal distribution $\mathcal{D}$ over an input space $\mathcal{X} \subseteq \mathbb{R}^p$ and a labeling function $f : \mathcal{X} \to \mathbb{R}$ that assigns a label Y (e.g. analyte concentration) to $\mathbf{x} \sim \mathcal{D}$, i.e. $\langle \mathcal{D}, f \rangle$. The *hypothesis* $h : \mathcal{X} \to \mathbb{R}$ is a function (i.e. the model) by which we try to approximate f using an appropriate learning algorithm (e.g. PLS regression). The standard setting in TL considers two domains: the source domain $\langle \mathcal{D}_S, f_S \rangle$ and the target domain $\langle \mathcal{D}_T, f_T \rangle$. In chemometrics, the source and target domains are commonly referred to as the primary and secondary "condition", respectively.

We will use $P(X)$, $P(Y|X)$ and $P(X, Y)$ to denote the marginal, conditional (Y conditioned on X) and joint probability distributions, respectively, and $\mathbb{E}[\cdot]$ to denote expectation. We will express the labeling function f as conditional probability distribution $P(Y|X)$ throughout the paper. By $P_{\mathrm{S}}(\cdot)$ and $P_{\mathrm{T}}(\cdot)$ we further denote distributions associated with source and target domains, respectively. The $n_{\mathrm{S}} \times p$ matrix $\mathbf{X}_{\mathrm{S}}$ (n_{S} samples by p variables) and $n_{\mathrm{T}} \times p$ matrix $\mathbf{X}_{\mathrm{T}}$ will denote source and target domain spectra and $\mathbf{y}_{\mathrm{S}}$ ($n_{\mathrm{S}} \times 1$) and $\mathbf{y}_{\mathrm{T}}$ ($n_{\mathrm{T}} \times 1$) the corresponding reference values, respectively. We further define the source means as $\boldsymbol{\mu}_{\mathrm{S}}^{\mathrm{x}} = (\mathbf{1}^{\mathrm{T}}\mathbf{X}_{\mathrm{S}})/n_{\mathrm{S}}$ and $\mu_{\mathrm{S}}^{\mathrm{y}} = (\mathbf{1}^{\mathrm{T}}\mathbf{y}_{\mathrm{S}})/n_{\mathrm{S}}$, and the target means as $\boldsymbol{\mu}_{\mathrm{T}}^{\mathrm{x}} = (\mathbf{1}^{\mathrm{T}}\mathbf{X}_{\mathrm{T}})/n_{\mathrm{T}}$ and $\mu_{\mathrm{T}}^{\mathrm{y}} = (\mathbf{1}^{\mathrm{T}}\mathbf{y}_{\mathrm{T}})/n_{\mathrm{T}}$.

2.2 Theory of learning from different domains

Statistical learning theory is based on the assumption that training and test data originate from the same domain, i.e. that they are sampled from a common, joint probability distribution $P(X, Y)$. Under this assumption, the training error from a (calibration) model is a good proxy for the test error [6]. However, in many real-world applications, this assumption does not hold and the distributions of training and test data might be considerably different.

In their seminal paper, entitled *A theory of learning from different domains*, Ben-David *et al.* investigated the conditions on the training (source domain) and test (target domain) distributions under which a classification model is expected to perform well [5]. The authors prove that the target error under a hypothesis h has the following upper bound

$$\epsilon_T(h) \;\leq\; \epsilon_S(h) \;+\; d_1(\mathcal{D}_S, \mathcal{D}_T) \;+\; \min\left[\mathbb{E}_{\mathcal{D}_S}[|f_S(\mathbf{x}) - f_T(\mathbf{x})|], \mathbb{E}_{\mathcal{D}_T}[|f_S(\mathbf{x}) - f_T(\mathbf{x})|]\right], \tag{1}$$

where $\epsilon_S(h)$ denotes the error in the source domain and the last term being a measure of the difference between the (unknown) labeling functions in the source and target domains. As we will see later, in some instances, it is reasonable to assume that this difference is small. The second term on the right-hand side of Eq. (1) is referred to as the L^1 divergence and is a measure of the difference between the marginal distributions of the source and target domains. Ben-David's theory formalizes the "intuition" that the source error is a good proxy for the target error if the domains are "similar" in terms of their marginal and conditional distributions $P(X)$ and $P(Y|X)$, respectively.

As we will see in the next sections, a widely used approach to TL is to find transformations (e.g. latent representations) of one (or both) domains in order to increase the similarity of the marginal distributions of source and target domain samples, i.e. to "make them look as if they were sampled from the same underlying distribution".

2.3 Nomenclature

We continue by introducing some terminology and definitions that are widely used in the TL community and that are relevant to applications in chemometrics.

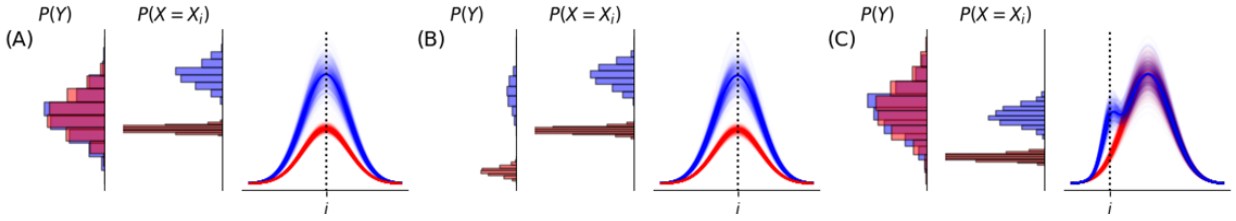

Figure 1: Types of domain shifts. A) Covariate shift. B) Prior shift. C) Conditional shift. The blue and red lines (spectra) correspond to the input data from a source and a target domain, respectively

Transfer learning and domain adaptation Transfer learning (TL) and domain adaptation (DA) are terms that are often used interchangeably. However, according to Pan et al. [1], TL relates more broadly to techniques that can cope with situations where the *source task* can be different from the *target task* (e.g. when a source model that discriminates between cars and trucks should be adapted to discriminate between bicycles and motorbikes), whereas in DA the task is usually the same across the domains. Along these lines, DA can be considered a special case of TL. Some authors use the terms *inductive* and *transductive* TL to refer to what we here call TL and DA, respectively. In contrast to research in computer vision, the vast majority of publications on TL in chemometrics and analytical chemistry have considered DA rather than TL so far.

Supervised, semi-supervised and unsupervised TL and DA Supervised TL/DA refers to the scenario when all training samples are labeled. We will refer to semi-supervised TL/DA, when some of the source training samples are unlabeled, and when some of the target training samples are unlabeled (e.g. when for some samples only the spectra are available but not the reference values). Unsupervised DA refers to situations, where all the target domain training samples are unlabeled. When the source and target tasks are different (i.e. when the labeling functions f_S and f_T introduced in the previous section are different), some subset of labeled samples from the target domain is required to adapt a source model to a target domain. As a result, unsupervised DA will not likely be feasible. However, when $f_S \approx f_T$, and if there exists a hypothesis h that performs well in both domains and the domains are sufficiently similar, unsupervised DA is likely to be feasible [7]. Unsupervised TL, on the other hand, refers to learning from only unlabelled data in both, the source and the target, domains and is considered out of scope for this contribution.

Covariate, prior and conditional shift We proceed by introducing some concepts that will be important to know when applying TL/DA in analytical chemistry. Figure 1 exemplifies the three ways how domains can differ (see also Table 1) using some simulated spectra. The red and blue lines correspond to input data from a source and a target domain, respectively and the histograms on the left show the probability density of $\mathbf{x_i} = [x_i^{(1)}, \ldots, x_i^{(n)}]^\intercal$ for the i-th variable/spectral channel and the response Y.

In case of *covariate shift*, the marginal distribution of the inputs $P(X)$ is different while the distribution of the response given the inputs $P(Y|X)$ (and thus also $P(Y)$) remains unchanged between the domains (Table 1). A typical scenario in analytical chemistry involving covariate shifts arises when samples of similar composition are measured on different spectrometers, where differences e.g. in the sensitivity of the detector can lead to a change of the width of the marginal distribution. The strength of the illumination source, on the other hand, will usually impact the amount of absorbance and thus affects the location of the input distribution. Together, these effects can lead to situations where $P_S(X) \neq P_T(X)$. Simple covariate shifts such as the one shown in Figure 1A can easily be accounted for by appropriately (re-) scaling the data (Figure 2A). However, more complicated changes in the covariance structure of the predictors eventually requires more sophisticated treatment. Generally speaking, covariate shifts can be corrected by employing transformations, e.g. of the form $W : \mathbb{R}^p \to \mathbb{R}^p$, either of the source (red), the target (blue) or both spectra such that $P_S(W_S(X)) \approx P_T(W_T(X))$ and $P_S(Y|W_S(X)) \approx P_T(Y|W_T(X))$ [8]. In fact, many preprocessing methods in chemometrics (e.g. baseline correction, scaling, standard normal variate transformation, piecewise direct standardization, multiplicative scatter correction, etc.) can be regarded as employing such transformations (Figure 3). As we will see later, covariate transformations that directly "align" the input distributions in a preprocessing step (section 3.1.2) or implicitly when modelling the response (section 3.2.1) are widely used in domain adaptation.

Prior shift refers to situations, where the distribution of the response $P(Y)$ is different in the domains (Figure 1B). Similar to the covariate shift example, the blue spectra have an overall higher intensity with the corresponding probability density being higher and more shallow for X_i, which however is due to smaller values and spread of the response Y. Using a (calibration) model derived from the blue spectra for inference on the red spectra might or might not give accurate results depending on whether the labeling function is different or not but is in general not advisable.

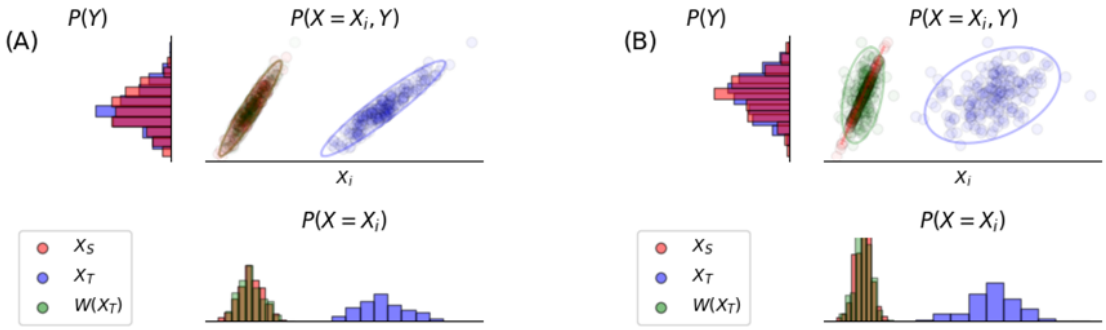

Figure 2: Covariate vs. conditional shift. The scatterplots show the empirical joint distribution of the i-th variable X_i and the response Y for the data shown in Figure 1A (covariate shift) (A) and Figure 1C (conditional shift) (B). The transformation $W(\cdot)$ re-scales the target domain samples such that $W(\mathbf{x}_T^{(k)}) = (\mathbf{x}_T^{(k)} - \boldsymbol{\mu}_{\text{T}}^{\text{x}})\frac{\sigma_S}{\sigma_T} + \boldsymbol{\mu}_{\text{S}}^{\text{x}}$, where σ_S and σ_T denote the standard deviation of the empirical distributions of $P_{\text{S}}(X)$ and $P_{\text{T}}(X)$, respectively. It can be seen that re-scaling of the predictors aligns the empirical joint distribution in case of covariate shift (A) but not when the difference between the domains is characterized by a conditional shift (B).

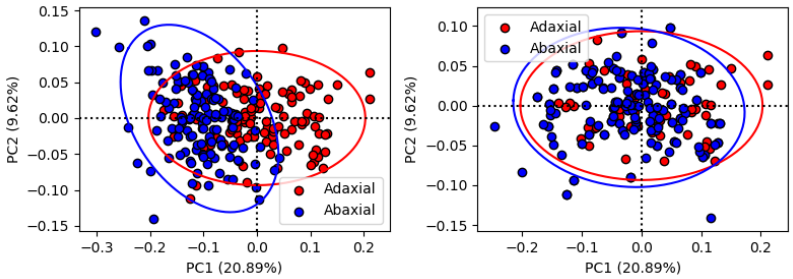

Figure 3: Effect of direct standardization (DS) on $P(X)$. Left: The first two principle components fitted to ATR-FTIR spectra of adaxial leaf surfaces of samples from the genus *Fragaria* (red) and projection of the spectra from abaxial leaf surfaces of the same samples (blue). Right: The same projection after direct standardization of the blue samples. Note that only a subset of the samples were used to derive the transformation matrix $\mathbf{F} = \mathbf{X}_{\text{blue}}^{\dagger}\mathbf{X}_{\text{red}}$.

Conditional shifts are the most frequent ones encountered in practice and occur, for instance, when an additional signal (e.g. from an interferrent) is introduced in one of the domains. In the example in Figure 1C, the blue spectra contain an additional peak, which similar to the covariate shift scenario, changes the marginal distribution of the inputs (i.e. $P_{\text{S}}(X) \neq P_{\text{T}}(X)$). In contrast to the covariate shift scenario from Figure 2A, where we could "align" the (empirical) joint distribution $P(X, Y)$ of the target to the one of the source domain by simply re-centering and scaling $\mathbf{X}_T$, the same transformation does not work with conditional shifts (Figure 2B). This is because the additional peak at X_i (blue spectra in Figure 1C) changes the conditional distribution $P(Y|X)$, i.e. the *correlation* between the i-th variable and the response. Without knowing the values for the response for (at least some of) the target domain spectra, we can not quantify that change. Thus, compensation of conditional shifts requires that labeled data (e.g. spectra with reference values) are obtained from the target domain in order to account for the additional variability not present in the source domain. However, as we will see in section 4, unsupervised compensation of the covariate shift can be sufficient if analyte and interferent signals are just weakly overlapping, i.e. if only a subset of the X-variables are affected by the conditional shift.

3 PREVIOUS WORK

The problems associated with domain shifts have been studied in analytical chemistry and chemometrics long before transfer learning have become a popular discipline in machine learning.

See [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37] for long-standing problems with domain shifts in chemometrics. More recent examples include calibration transfer (CT)

Table 1: Different ways of how source and target domains can differ in terms of the underlying distribution of the inputs $\boldsymbol{X}$, the response $\boldsymbol{Y}$ or the response conditioned on the inputs, i.e. $Y|X$.

Concept	Formal Definition	Description
Covariate shift	$P_{\text{S}}(X) \neq P_{\text{T}}(X)$ $P_{\text{S}}(Y\|X) \approx P_{\text{T}}(Y\|X)$	Change in the covariance structure of the predictors and/or spectral offset, e.g. due to baseline effects, peak broadening etc.
Prior shift	$P_{\text{S}}(Y) \neq P_{\text{T}}(Y)$	Change in the distributon of the response, e.g. change in concentration range of the analyte of interest.
Conditional shift	$P_{\text{S}}(X) \neq P_{\text{T}}(X)$ $P_{\text{S}}(Y\|X) \neq P_{\text{T}}(Y\|X)$	Change in the correlation between predictors and response, e.g. due to presence of a new interferrent.

problems, where the aim is to transfer calibrations between similar devices (typically spectrometers) with (slightly) different instrumental responses [38, 39], applications where the physical and/or chemical properties of the test samples differ from those of the calibration samples [40, 41, 42] or some environmental condition (e.g. temperature, humidity) has changed between the time of calibration and application of the model [43]. In general, a source (or primary) calibration model may not be maintained to new conditions, e.g. due to instrumental drift or uncalibrated spectral features appearing in new target (or secondary) samples occurring later in time. As a result, calibration transfer and/or maintenance (CTM) mechanisms are needed to accommodate the new chemical, physical, environmental, and/or instrumental effects not spanning the source (calibration) domain. To be consistent with terminology introduced in prior sections, we will use the TL/DA-based domain nomenclature (i.e. *source* and *target*) as a pseudonym for the chemometrics-based domain nomenclature (i.e. primary and secondary). Also, we will use the term *domain difference* as an umbrella term for the unmodeled sources of spectral variability and/or response variability (prior shift) within the target samples that are not accounted for by the source calibration model. The vast majority of CTM approaches roughly fall into four categories:

1. **Adjust the instruments**. For example, one can use libraries of reference samples measured on both the source and target devices to adjust the instrument response or wavelength registry of the target device to match the instrument response of the source device.
2. **Adjust the spectra**. One can preprocess the spectra from the source and the target conditions using techniques such as e.g. baseline and multiplicative scatter correction (MSC), finite impulse response (FIR) filters, derivatives, wavelets, and/or wavelength selection to obtain a transformed set of spectra that are robust to domain differences. Procrustes analysis-based approaches (e.g, piecewise direct standardization; PDS) that aim at making the target spectra match or "look" like the spectra from the source condition via rotation, dilation, and translation fall in this category, too.
3. **Adjust the model**. Approaches from this category update (or rebuild) the source calibration model by augmenting the original coefficient matrices involving source spectra and reference measurements with auxiliary matrices involving both source and target domain information. The idea is to infuse the original calibration model with *a priori* chemical and/or spectroscopic information such that the modified calibration model is less sensitive to domain differences.
4. **Adjust the predictions**. For example, slope and bias correction can be used to adjust the predictions of a source model when applied to target domain samples.

In this section, we review a select set of CTM approaches from the **second** and **third** category that relate to the concepts in Table 1 and can thus be viewed from a domain adaptation perspective. As we shall see later on, the vast majority of CTM approaches address covariate and conditional shifts, while CTM approaches that can accommodate prior shifts are less well represented.

3.1 Adjust the Spectra

The CTM techniques addressed here focus on linearly transforming the target and/or source spectra such that both sets of spectra are subsequently shape matched with respect to domain distributions. Before transformation, the source and target spectra are typically characterized by covariate ($P_{\text{S}}(X) \neq P_{\text{T}}(X)$) and/or conditional shifts ($P_{\text{S}}(Y|X) \neq P_{\text{T}}(Y|X)$). After transformation, the intent is to match (as much as possible) the domain distributions of the source and target samples such that ($P_{\text{S}}(X) \approx P_{\text{T}}(X)$). Once the linear transformation matrix $\mathbf{W}$ has been computed, all

subsequent spectra will be projected via $\mathbf{X}_{\text{PROJ}} = \mathbf{XW}$ and calibration and prediction will be done in the transformed space.

3.1.1 Direct Standardization and Variants

In chemometrics, perhaps the most popular technique in this CTM category is direct standardization (DS) and its variants [12, 36, 37, 44]. In DS, one maps the target spectra onto the source spectra by a linear transformation matrix $\mathbf{W}$ that minimizes $||\mathbf{X}_{\text{T}}\mathbf{W} - \mathbf{X}_{\text{S}}||_2$ on a set of matched samples (i.e. calibration standards). Several variants of DS exist, e.g. piecewise direct standardization (or PDS) or penalized/regularized versions that encourage certain properties such as smoothness and/or sparsity [44]. All these methods have been developed based on the intuition that matched samples should return equal (spectroscopic) signals when measured on different devices. However, (successful) instrument standardization usually has the (side) effect of aligning the marginal distributions of source and target domain samples and thus correcting for covariate shift (Figure 3). Similar observations can be made with other Procustes analysis-type methods like Standard Normal Variate (SNV) transformation that utilize translation (mean-centering), rotation and stretching [21] or other preprocessing methods such as MSC or baseline correction. Thus, some of the most widely used (preprocessing) techniques in chemometrics actually have a DA interpretation. As we will see in the next section, several DA methods operate at the distributional rather than at the sample level in order to find an appropriate transformation $\mathbf{W}$.

3.1.2 Generalized Eigenproblems

Several of the earlier DA methods that have been adopted in chemometrics solve so-called Generalized Eigenproblems (GE) of the type $\mathbf{Bw} = \lambda\mathbf{Cw}$ involving two matrices $\mathbf{B}$ and $\mathbf{C}$ of size $p \times p$ [34]. A GE often emerges from the maximization of a Rayleigh quotient, which can be subsequently transformed into a GE via traditional calculus-based Lagrangian multiplier approaches:

$$\underbrace{\max_{\mathbf{w}} \frac{\mathbf{w}^{\text{T}}\mathbf{Bw}}{\mathbf{w}^{\text{T}}\mathbf{Cw}}}_{\text{Rayleigh quotient}} \Rightarrow \max_{\substack{\mathbf{w} \\ \mathbf{w}^{\text{T}}\mathbf{Cw}=1}} (\mathbf{w}^{\text{T}}\mathbf{Bw}) \Rightarrow \max_{\mathbf{w}} (\mathbf{w}^{\text{T}}\mathbf{Bw} - \lambda(\mathbf{w}^{\text{T}}\mathbf{Cw} - 1)) \Rightarrow \mathbf{Bw} - \lambda\mathbf{Cw} = \mathbf{0} \Rightarrow \underbrace{\mathbf{Bw} = \lambda\mathbf{Cw}.}_{\text{GE}} \quad (2)$$

We will be interested in the largest k ($k \ll p$) eigenvalue-eigenvector solution pairs $\{\lambda_i, \mathbf{w}_i\}$, $i = 1, 2, ..., k$, whereby the samples (e.g. spectra) $\mathbf{X}$ are projected onto a lower k-dimensional subspace such that $\mathbf{X}_{\text{PROJ}} = \mathbf{XW}$ with $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$. In the examples to follow, $\mathbf{B}$ and $\mathbf{C}$ will encode total and between-domain scatter information, respectively (or vice versa) such that solving Eq. (2) yields subspaces that explain a large amount of the variation in $\mathbf{X}$ while being (nearly) invariant with respect to domain differences (see Figure 4). We now examine two classes of GEs which both involve Laplacian matrices.

Distance-Based Laplacians A Laplacian matrix $\mathbf{L}$ can be used to describe the relationship between (spectral) samples i and j in a graph. The matrix $\mathbf{L}$ is a symmetric, $n \times n$ (i.e. samples times samples) matrix characterized by having zero sums across all its rows (and columns) and being positive-semidefinite (i.e. all of its eigenvalues are positive). The Laplacian matrix has many properties but perhaps the most interesting one for CTM purposes is its relationship with Principal Component Analysis (PCA) [45]. Suppose $\mathbf{u} = \mathbf{Xw}$ with $\mathbf{w}$ being a $p \times 1$ (latent variable) vector and $\mathbf{X} = [\mathbf{X}_{\text{S}}; \mathbf{X}_{\text{T}}]$ is a $n \times p$ matrix consisting of both source and target spectra that has been mean-centered, then it can be shown that [38]

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{x}_i^{\text{T}}\mathbf{w} - \mathbf{x}_j^{\text{T}}\mathbf{w})^2\ \mathbf{A}_{i,j} \\
&= \mathbf{w}^{\text{T}}\left[\sum_{i=1}^{n}\mathbf{x}_i\mathbf{D}_{i,i}\mathbf{x}_i^{\text{T}} - \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{x}_i\mathbf{A}_{i,j}\mathbf{x}_j^{\text{T}}\right]\mathbf{w} \\
&= \mathbf{w}^{\text{T}}\mathbf{X}^{\text{T}}(\mathbf{D} - \mathbf{A})\mathbf{Xw} \\
&= \mathbf{u}^{\text{T}}\mathbf{Lu}.
\end{aligned} \quad (3)$$

$\mathbf{A}$ and $\mathbf{D}$ denote the adjacency and degree matrix of a simple, undirected graph defining the weights between, and connectedness of, each of the n data points, respectively and $\mathbf{L} = \mathbf{D} - \mathbf{A}$. If $\mathbf{A}_{i,j} = 1\ \forall\ i, j$ the vector $\mathbf{u}$ that maximizes Eq. (3) corresponds to the (unscaled) loading vector of the first principle component. However, one may prefer to impose some weights $\mathbf{A}_{i,j}$ to achieve various goals such as enlarging the distance between samples belonging

to different classes or minimizing the distance between domains (Figure 4). Hence, the Laplacian matrix allows one to incorporate *a priori* knowledge about the relationship between samples or classes of samples.

In CTM applications (and DA), the goal is to find (domain-invariant) data representations where the differences between domains are small. To this end, we can formulate a GE with $\mathbf{B} = \mathbf{X}^{\mathrm{T}}\mathbf{L}\mathbf{X}$ and $\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ using a Laplacian matrix of the type [45]

$$\mathbf{L}_{ij} = \begin{cases} \sum_{j=1}^{p} d_{\mathrm{ij}}^{\mathrm{lab}}, & i = j \\ -d_{\mathrm{ij}}^{\mathrm{lab}}, & i \neq j \end{cases} \quad \text{where } d_{\mathrm{ij}}^{\mathrm{lab}} = \begin{cases} 0, & \text{samples } i \text{ and } j \text{ belong to different domains} \\ d_{\mathrm{ij}}, & \text{otherwise.} \end{cases} \tag{4}$$

with d_{ij} denoting the reciprocal Euclidean distance between data points i and j. Note that a slight modification to $d_{\mathrm{ij}}^{\mathrm{lab}}$ can instead yield representations where differences between domains are large:

$$d_{\mathrm{ij}}^{\mathrm{lab}} = \begin{cases} 0, & \text{samples } i \text{ and } j \text{ belong to the same domains} \\ d_{\mathrm{ij}}, & \text{otherwise.} \end{cases} \tag{5}$$

An example for these types of GEs—domain minimization and domain maximizaton—are shown in the middle and right-most plots, respectively, of Figure 4 for a NIR data set containing spectra from four different tree species[1]. In Section 4, we will illustrate how this Laplacian-based GE can be used for CTM.

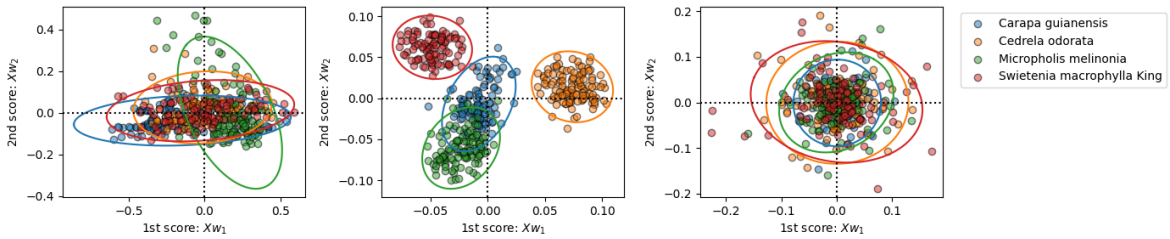

Figure 4: Laplacian projection schemes. Three different two-dimensional projections of the NIR spectra of four species of wood samples are shown. The left figure shows the ordinary PCA projection while the next two show Laplacian-based projection schemes. The middle plot shows a domain-separation projection scheme based upon Eq.(5) while the right plot shows the domain-minimization projection scheme based upon Eq.(4).

Transfer Component Analysis and Scatter Component Analysis Two related projection schemes from the TL community, i.e. Transfer Component Analysis (TCA) [46] and Scatter Component Analysis (SCA) [47], similarly leverage a Laplacian matrix but instead aim to maximize the following quotient: (total scatter)/(domain scatter). Maximizing the numerator aims to preserve the total variability of the data while minimizing the denominator encourages finding a representation for which the source and target domains are similar. TCA expresses the (total scatter)/(domain scatter) trade-off as $\mathbf{B} = \mathbf{K}\mathbf{H}\mathbf{K}$ and $\mathbf{C} = \mathbf{I} + \delta\mathbf{K}\mathbf{L}\mathbf{K}$ in Eq. (2) with $\mathbf{K}$ being some *kernel* of $\mathbf{X} = [\mathbf{X}_{\mathrm{s}}; \mathbf{X}_{\mathrm{T}}]$ (in case of a linear kernel $\mathbf{K} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$), $\mathbf{H} = \mathbf{I} - \frac{1}{n_{\mathrm{S}}+n_{\mathrm{T}}}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ (i.e. the centering matrix for $\mathbf{K}$) and $\mathbf{L}$ being defined as

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}^{11} & \mathbf{L}^{12} \\ \mathbf{L}^{21} & \mathbf{L}^{22} \end{bmatrix} \mathbf{L}^{11} = \frac{1}{n_{\mathrm{s}}^2}\mathbf{1}_{(n_{\mathrm{S}},n_{\mathrm{S}})},\ \mathbf{L}^{22} = \frac{1}{n_{\mathrm{T}}^2}\mathbf{1}_{(n_{\mathrm{T}},n_{\mathrm{T}})},\ \mathbf{L}^{12} = -\frac{1}{n_{\mathrm{s}}n_{\mathrm{T}}}\mathbf{1}_{(n_{\mathrm{S}},n_{\mathrm{T}})},\ \mathbf{L}^{21} = -\frac{1}{n_{\mathrm{s}}n_{\mathrm{T}}}\mathbf{1}_{(n_{\mathrm{T}},n_{\mathrm{S}})}. \tag{6}$$

Here, $\mathbf{1}_{(n_{\mathrm{S}},n_{\mathrm{T}})}$ indicates a matrix of all ones of size $n_{\mathrm{s}} \times n_{\mathrm{T}}$. It can be shown that the corresponding GE minimizes the so-called *Maximum Mean Discrepancy*

$$\mathrm{MMD}(\mathbf{X}_{\mathrm{s}}, \mathbf{X}_{\mathrm{T}}) = \left\| \frac{1}{n_S}\sum_{i=1}^{n_S}\phi(\mathbf{x}_{S_i}) - \frac{1}{n_T}\sum_{i=1}^{n_T}\phi(\mathbf{x}_{T_i}) \right\|_{\mathcal{H}}, \tag{7}$$

where $\phi(\cdot)$ denotes the feature map associated with the kernel and $\|\cdot\|_{\mathcal{H}}$ denotes the reproducing kernel Hilbert space (RKHS). For a linear kernel, the MMD corresponds to the difference between the means (i.e. first order moments) of source and target distributions, whereas higher order moments (e.g. covariances) will be aligned when using polynomial or Gaussian kernels [48]. In addition to total and domain scatter, SCA also aims to preserve between-class information (in a classification context). As outlined by Andries in [34], SCA is equivalent to TCA except that $\mathbf{L}$ is replaced with a regularized version of itself: $\mathbf{L} := \mathbf{L} + \delta\mathbf{I}$, when no class structure is present in the data. For both TCA and SCA, the collection of eigenvectors $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$ and eigenvalues $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ associated with the largest

[1] http://groupwentzell.chemistry.dal.ca/software.html, accessed 04/25/2021

eigenvalues yield the desired projection: $\mathbf{K}_{\text{PROJ}} = \mathbf{K}\mathbf{W}\mathbf{\Lambda}^{-1/2}$ (note that when aiming to project a test sample, its kernel with respect to the calibration set must be computed first). With respect to CTM applications, SCA and TCA were explored in [34, 49], and they were deemed to be not that effective. In our view the main reason for the limited success is the fact that these methods align (source and target) distributions effectively only in the non-linear case (i.e. when using non-linear kernels). This, however, is prohibitive for small samples and/or if the relationship between inputs (e.g. spectra) and the response (e.g. concentration) is approximately linear [49, 50].

3.1.3 Orthogonal Projections

Othogonal Projection (OP) methods, as a whole, strive to make a calibration model insensitive to interfering sources of variation not present in the calibration set—see [51, 52, 53] and the references therein. Note that we also want to distinguish between OP methods and orthogonal signal correction (OSC) methods that have been proposed for removing detrimental information orthogonal to $\mathbf{y}$ from $\mathbf{X}$–see [53] for a thorough taxonomy of OSC methods. While there are many OP variants, we will examine one of the simplest, i.e., transfer orthognal projection (TOP), since it addresses directly the TL/DA problem [51].

Transfer Orthogonal Projection (TOP) In TOP, one collects a matched set of samples. Instead of working with these spectra directly, one uses difference spectra. The difference spectra are computed as the difference between the signals of the same samples measured under different conditions (e.g. using different spectrometers). As a result, the corresponding reference value should be (close to) zero (eliminating the need to use a reference method to obtain a reference value).

Let $\mathbf{G}$ denote the $m \times p$ matrix of difference spectra (i.e. $\mathbf{G} = \mathbf{X}_{\text{S}}^{(\text{MS})} - \mathbf{X}_{\text{T}}^{(\text{MS})}$), the matrix containing the unwanted sources of spectral variation (e.g. between-instrument variation). If we perform Principal Component Analysis (PCA, or the SVD) on the mean-centered version of $\mathbf{G}$ such that $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\text{T}}$, we can define two projections: projections onto span$(\mathbf{G})$ via $\mathbf{P}_{\text{G}} = \mathbf{G}^{+}\mathbf{G} = \mathbf{V}\mathbf{V}^{\text{T}}$ and projections onto its orthogonal complement via $\mathbf{P}_{\text{G}}^{\perp} = \mathbf{I} - \mathbf{P}_{\text{G}}$. Note that a key parameter in TOP is the number of loading vectors in $\mathbf{V}$ to use: let $1 \leq k < m$ be the number of loading vectors such that we replace $\mathbf{V}$ with $\mathbf{V}_{\text{k}} = [\mathbf{v}_1, \ldots, \mathbf{v}_{\text{k}}]^{\text{T}}$. Hence, the linear transformation matrix that we seek is $\mathbf{W} = \mathbf{P}_{\text{G}}^{\perp}$ whereby $\mathbf{X}_{\text{PROJ}} = \mathbf{X}\mathbf{W}$. The goal of the null projection is to remove k dimensions that best capture between-instrument variability. If k is too small, then not all unwanted inter-instrument variances are removed, and the correction is not complete. If k is too big, then too much variance is removed. One major pitfall of TOP (and of OP methods in general) is that there is no guarantee that the spectra contained in $\mathbf{G}$ do not contain meaningful reference value information, e.g., there could be (and there often is) a large amount of spectral overlap between $\mathbf{G}$ and the analyte signal of interest. As a result, OP methods can be quite deleterious if not used appropriately.

3.1.4 Generalized Least Squares

As noted in the discussion regarding TOP in Section 3.1.3, the spectra in $\mathbf{G}$ contains *non-analyte* domain-difference information that we want to desensitize the calibration model against. However, this desensitizing operation can be carried out instead by a covariance-based "pre-whitening" operation where the source spectra $\mathbf{X}_{\text{s}}$ is post-multiplied by a matrix $\mathbf{W}$ such that $\mathbf{X}_{\text{PROJ}} = \mathbf{X}_{\text{S}}\mathbf{W}$. This is the stated goal of Generalized Least Squares (GLS) [27]. Unlike TOP which projects away from span$(\mathbf{G})$, GLS instead shrinks the spectra in directions that are dominated by inter-instrument variance. GLS first computes the difference between mean-centered matched spectra associated with the source ans target spectra (denoted as $\mathbf{X}_{\text{S}}^{(\text{MS})}$ and $\mathbf{X}_{\text{T}}^{(\text{MS})}$). The difference spectra is defined as the matrix $\mathbf{L} = (\mathbf{X}_{\text{S}}^{(\text{MS})} - \mathbf{1}\boldsymbol{\mu}_{\text{S}}^{(\text{MS})}) - (\mathbf{X}_{\text{T}}^{(\text{MS})} - \mathbf{1}\boldsymbol{\mu}_{\text{T}}^{(\text{MS})})$ where $\boldsymbol{\mu}_{\text{T}}^{(\text{MS})}$ and $\boldsymbol{\mu}_{\text{S}}^{(\text{MS})}$ are the mean spectra associated with $\mathbf{X}_{\text{T}}^{(\text{MS})}$ and $\mathbf{X}_{\text{S}}^{(\text{MS})}$, respectively. (Again, the domain differences captured by $\mathbf{L}$ may not necessarily include only inter-instrument variation.) We then compute a re-scaled covariance matrix and rewrite it via its SVD: $\mathbf{C} = \mathbf{L}^{\text{T}}\mathbf{L} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^{\text{T}}$. We then replace $\mathbf{\Sigma}^2$ with another diagonal matrix $\mathbf{F} = \text{diag}(f_1, f_2, \ldots, f_p)$ whose components are defined by $f_i = \sqrt{\gamma^2/(s_i^2 + \gamma^2)}$, $\gamma > 0$. The linear transformation (and pre-whitening) matrix that we seek is expressed as $\mathbf{W} = \mathbf{V}\mathbf{F}\mathbf{V}^{\text{T}} = \sum_{i=1}^{p} f_i \mathbf{v}_i \mathbf{v}_i^{\text{T}}$. The diagonal element $0 \leq f_i < 1$ filters (or shrinks) the contribution of the i^{th} loading vector: when γ is small ($\gamma \ll s_i$), the contribution of $\mathbf{v}_i$ is negligible since $f_i \approx 0$; when γ is large ($\gamma \gg s_i$), then the contribution of $\mathbf{v}_i$ is left intact since $f_i \approx 1$. An excellent mathematical description of GLS for drift correction can be found in [31].

3.2 Adjust the Model

In the previous section we have reviewed a select set of CTM approaches that aim at compensating for covariate shift between source and target domain explicitly through some sort of preprocessing of the input data. In what follows we will briefly review some important approaches from the chemometrics field that implicitly compensate for domain differences while fitting the response. The approaches outlined in this subsection address covariate shifts

($P_s(X) \neq P_T(X)$) and/or conditional shifts ($P_s(Y|X) \neq P_T(Y|X)$). In some instances (see Section 3.2.1), target shifts can also be addressed.

3.2.1 Generalized Tikhonov Regularization

Generalized Tikhonov Regularization (GTR) can accommodate many different types of constraints. The augmented system of linear equations, and its corresponding least squares minimization problem

$$\left[\begin{array}{c} \mathbf{X}_s \\ \tau\mathbf{G} \end{array} \right] \mathbf{b} = \left[\begin{array}{c} \mathbf{y}_s \\ \tau\mathbf{h} \end{array} \right] \;\Leftrightarrow\; \min_{\mathbf{b}} ||\mathbf{X}_s\mathbf{b} - \mathbf{y}_s||_2^2 + \tau^2 ||\mathbf{G}\mathbf{b} - \mathbf{h}||_2^2 \tag{8}$$

characterizes the overall framework [54, 55, 56, 57]. Unless indicated otherwise, the source spectra $\mathbf{X}_s$ and source reference measurements $\mathbf{y}_s$ are mean-centered with respect to their source means: $\mathbf{X}_s := \mathbf{X}_s - \mathbf{1}\boldsymbol{\mu}_s^x$ and $\mathbf{y}_s := \mathbf{y}_s - \mathbf{1}\mu_s^y$. The matrix $\mathbf{G}$ and vector $\mathbf{h}$ may or may not be mean-centered, depending on the context. The penalized least squares framework of Eq. (8) approximates the following: solve $\mathbf{X}_s\mathbf{b} = \mathbf{y}_s$ subject to $\mathbf{G}\mathbf{b} = \mathbf{h}$. The larger the value of the non-negative regularization parameter τ, the more important the equality constraint $\mathbf{G}\mathbf{b} = \mathbf{h}$ is relative to satisfying the model fit of the source samples. When $\tau = 0$, there is no calibration updating. Historically, the matrix $\mathbf{G}$ and the vector $\mathbf{h}$ encode prior knowledge with respect to mathematical and statistical concerns: smoothness, monotonicity or piecewise linearity. For CTM purposes, however, $\mathbf{G}$ and $\mathbf{h}$ encode prior knowledge with respect to domain differences. We will look at many different forms of the matrix-vector pair $(\mathbf{G}, \mathbf{h})$, and each form corresponds to a qualitatively different CTM mechanism.

The most common GTR case occurs when $\mathbf{h} = \mathbf{0}$, and this is also of great interest from the CTM perspective. In this case, Eq. (8) and its solution can be expressed as

$$\left[\begin{array}{c} \mathbf{X}_S \\ \tau\mathbf{G} \end{array} \right] \mathbf{b} = \left[\begin{array}{c} \mathbf{y}_S \\ \mathbf{0} \end{array} \right] \;\Leftrightarrow\; \min_{\mathbf{b}} \; ||\mathbf{X}_S\mathbf{b} - \mathbf{y}_S||_2^2 + \tau^2 ||\mathbf{G}\mathbf{b}||_2^2 \;\Leftrightarrow\; \mathbf{b} = \left(\mathbf{X}_S^T\mathbf{X}_S + \tau^2\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{X}_S^T\mathbf{y}_S. \tag{9}$$

This framework has many interpretations. Statistically, the presence of the second term $\tau^2\mathbf{b}^T(\mathbf{G}^T\mathbf{G})\mathbf{b}$ in Eq.(9) has a covariate shift interpretation: An acknowledgment that the distributions of the source and target domains are different (in terms of covariance), and that a correction is required. Geometrically, the corresponding augmented linear system in Eq.(9) has an oblique projection interpretation: as τ increases, the regression vector $\mathbf{b}$ is increasingly becoming perpendicular to (or being projected away from) the subspace spanned by $\mathbf{G}$. Eq. (9) also states that $\mathbf{G}\mathbf{b} = \mathbf{0}$, i.e. that $\mathbf{b}$ is perpendicular to $\mathbf{G}$. Spectroscopically, this implies that the span$(\mathbf{G})$ is a subspace spanned by spectral interferents containing no analyte information. Hence, the domain differences embodied by $\mathbf{G}$ can be interpreted as spectral interferents that we want to desensitize the calibration model against.

There is an often-neglected relationship between between GTR and the CTM approaches of Section 3.1 where we pre-process the spectra via $\mathbf{X}_{\text{PROJ}} = \mathbf{X}\mathbf{W}$ before model building. In Eq. (9), if $\mathbf{G}$ is invertible, we can transform *Generalized* Tikhonov Regularization to *Standardized* Tikhonov Regularization (STR, or ordinary ridge regression) [56]:

$$\min_{\mathbf{b}} \; ||\mathbf{X}_s\mathbf{b} - \mathbf{y}_s||_2^2 + \tau^2 ||\mathbf{G}\mathbf{b}||_2^2 \;=\; \min_{\boldsymbol{\beta}} \; \left|\left|\left(\mathbf{X}_s\mathbf{G}^{-1}\right)\boldsymbol{\beta} - \mathbf{y}_s\right|\right|_2^2 + \tau^2 ||\boldsymbol{\beta}||_2^2 \text{ where } \boldsymbol{\beta} = \mathbf{G}\mathbf{b}. \tag{10}$$

Note that in Eq. (10), the solution $\boldsymbol{\beta}$ of the linear system $\left(\mathbf{X}_s\mathbf{G}^{-1}\right)\boldsymbol{\beta} = \mathbf{y}_s$ is being obtained via ridge regression but other regression techniques such as PLS could be used. If we set $\mathbf{W} = \mathbf{G}^{-1}$ where $\mathbf{X}_{\text{PROJ}} = \mathbf{X}\mathbf{W} = \mathbf{X}\mathbf{G}^{-1}$, then one performs regression (and make predictions) on spectra that already has been pre-processed (the "beta" space) and the outcome will be the same as the one obtained by GTR. If $\mathbf{x}_{\text{new}}$ is a novel spectrum that has been appropriately mean-centered, then its prediction y_{new} will be

$$\text{STR prediction} = y_{\text{new}} = (\mathbf{x}_{\text{new}}\mathbf{G}^{-1})\boldsymbol{\beta} = (\mathbf{x}_{\text{new}}\mathbf{G}^{-1})(\mathbf{G}\mathbf{b}) = \mathbf{x}_{\text{new}}\mathbf{b} = \text{GTR prediction}. \tag{11}$$

Even if $\mathbf{G}$ is not invertible, we can borrow an approach from [58] whereby we replace $\mathbf{G}$ with a new penalty matrix $\mathbf{H}$ in Eqs. (9)(10): $\mathbf{H} = (1-\alpha)\mathbf{P}_G + \alpha\mathbf{P}_G^{\perp}$. The matrix $\mathbf{H}$ has the nice property that its inverse is easy to compute: $\mathbf{H}^{-1} = (1-\alpha)^{-1}\mathbf{P}_G + \alpha^{-1}\mathbf{P}_G^{\perp}$. (Recall that $\mathbf{P}_G\mathbf{b} = \mathbf{0}$ projects $\mathbf{b}$ away from span$(\mathbf{G})$ while $\mathbf{P}_G^{\perp}\mathbf{b} = \mathbf{0}$ projects $\mathbf{b}$ toward span$(\mathbf{G})$.) Hence, in the orthogonality constraint $\mathbf{H}\mathbf{b} = \mathbf{0}$ in Eq.(9), as α goes from 0 to 1, the effect is to move the model vector $\mathbf{b}$ toward span$(\mathbf{G})$. For CTM purposes, the value of α should be chosen close to zero, i.e., projecting $\mathbf{b}$ away from the undesirable domain difference information embodied by span$(\mathbf{G})$.

Matched Differences A matched set of samples across the source and target domains ($\mathbf{X}_S^{(MS)}$ and $\mathbf{X}_T^{(MS)}$) has, in principle, the same reference values across instruments (or measurement conditions). As a result, the set of matched differences correspond to non-analyte spectra in Eq. (8): $\mathbf{G} = \mathbf{X}_S^{(MS)} - \mathbf{X}_T^{(MS)}$ and $\mathbf{h} = \mathbf{0}$. Spectroscopically, the matrix $\mathbf{G}$ now

reflects instrument-to-instrument differences that we want our calibration model to be insensitive to, and in a GTR context, span($\mathbf{G}$) geometrically defines an undesirable subspace that we want the model vector $\mathbf{b}$ to point away from [11, 59, 30]. GTR with matched differences in $\mathbf{G}$ invokes a covariate shift correction approach in Table 1. Difference spectra has a long history in CTM applications—see [60, 61, 30] and references therein.

CTM approaches using matched samples can also seemingly look quite different from GTR but are in fact equivalent. One such approach is the graph regularization scheme proposed by Nikzad-Langerodi and Sobieczky in [38], which employs a penalty term based upon the Laplacian of a special graph, where only matched samples are connected by an edge, in order to reduce inter-device variance. This idea can be cast into the GTR framework as follows:

$$\min_{\mathbf{b}} ||\mathbf{X}_{\text{S}}\mathbf{b} - \mathbf{y}_{\text{S}}||_2^2 + \tau \mathbf{b}^{\text{T}}\mathbf{\Lambda}\mathbf{b} \quad \text{where} \quad \mathbf{\Lambda} = \begin{bmatrix} \mathbf{X}_{\text{S}}^{(\text{MS})} \\ \mathbf{X}_{\text{T}}^{(\text{MS})} \end{bmatrix}^{\text{T}} \begin{bmatrix} \mathbf{I} & -\mathbf{I} \\ -\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{\text{S}}^{(\text{MS})} \\ \mathbf{X}_{\text{T}}^{(\text{MS})} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{\text{S}}^{(\text{MS})} \\ \mathbf{X}_{\text{T}}^{(\text{MS})} \end{bmatrix}^{\text{T}} \begin{bmatrix} \mathbf{X}_{\text{S}}^{(\text{MS})} - \mathbf{X}_{\text{T}}^{(\text{MS})} \\ -\mathbf{X}_{\text{S}}^{(\text{MS})} + \mathbf{X}_{\text{T}}^{(\text{MS})} \end{bmatrix} \tag{12}$$

Setting $\mathbf{G}$ to be the difference between matched samples, i.e., $\mathbf{G} = \mathbf{X}_{\text{S}}^{(\text{MS})} - \mathbf{X}_{\text{T}}^{(\text{MS})}$, the matrix $\mathbf{\Lambda}$ can be re-written as:

$$\mathbf{\Lambda} = (\mathbf{X}_{\text{S}}^{(\text{MS})})^{\text{T}}\mathbf{G} + (\mathbf{X}_{\text{T}}^{(\text{MS})})^{\text{T}}(-\mathbf{G}) = (\mathbf{X}_{\text{S}}^{(\text{MS})} - \mathbf{X}_{\text{T}}^{(\text{MS})})^{\text{T}}\mathbf{G} = \mathbf{G}^{\text{T}}\mathbf{G}. \tag{13}$$

Hence, the penalty term is the same as the GTR penalty term in Eq. (9), i.e. , $\tau \mathbf{b}^{\text{T}}\mathbf{L}\mathbf{b} = \tau \mathbf{b}^{\text{T}}(\mathbf{G}^{\text{T}}\mathbf{G})\,\mathbf{b} = \tau\,||\mathbf{G}\mathbf{b}||_2^2$.

First- and Second-Order Moment Differences Matched samples and/or labeled target samples should be used whenever they are available. However, the acquisition of such samples is often not feasible. For instance, matched samples are not available for problems where the sample matrix changes between the domains. On the other hand, acquiring unlabeled target samples that are not matched are in general easier to obtain with respect to expense and time. Suppose we only have two sets of calibration samples: labeled source samples $(\mathbf{X}_{\text{S}}, \mathbf{y}_{\text{S}})$ and unlabeled target samples $\mathbf{X}_{\text{T}}$, which is all we need to correct for covariate shifts given that $\mathbf{X}_{\text{T}}$ is a representative sample from the target domain (see Figure 2A). In this case, two mechanisms that capture domain differences between source and target domain naturally arise: Differences between first (i.e. means) and second order moments (i.e. covariances).

Simply subtracting the mean source spectrum from the mean target spectrum yields $\mathbf{G} = \boldsymbol{\mu}_{\text{T}}^{\text{x}} - \boldsymbol{\mu}_{\text{S}}^{\text{x}}$ and has been used in [62, 34]. Since the matrix $\mathbf{G}$ involves means, the GTR penalty term $||\mathbf{G}\mathbf{b}||_2 = ||(\boldsymbol{\mu}_{\text{T}}^{\text{x}} - \boldsymbol{\mu}_{\text{S}}^{\text{x}})\mathbf{b}||_2$ can be considered a *first moment* update. Here, one tries to make the calibration model indifferent to domain differences, as characterized by a single difference spectrum. Such a simple characterization will not likely suffice in capturing meaningful scatter differences across instruments or conditions. To better accommodate covariate shifts in CTM applications, a penalty term involving second order moment-differences can be considered and this was first proposed in [40] for PLS type models. Employing the GTR framework, this amounts to:

$$\min_{\mathbf{b}} ||\mathbf{X}_{\text{S}}\mathbf{b} - \mathbf{y}_{\text{S}}||_2^2 + \tau^2\,||\mathbf{G}\mathbf{b}||_2^2 \quad \text{where} \quad \mathbf{G} = \left|\mathbf{C}_{\text{S}} - \mathbf{C}_{\text{T}}\right|, \quad \mathbf{C}_{\text{S}} = \frac{1}{n_{\text{S}}}\mathbf{X}_{\text{S}}^{\text{T}}\mathbf{X}_{\text{S}} \text{ and } \mathbf{C}_{\text{T}} = \frac{1}{n_{\text{T}}}\mathbf{X}_{\text{T}}^{\text{T}}\mathbf{X}_{\text{T}}. \tag{14}$$

To account for first order moment-differences, both $\mathbf{X}_{\text{S}}$ and $\mathbf{X}_{\text{T}}$ are locally mean-centered with respect to their own domain means. It is important to acknowledge that in general, the component-wise covariance difference matrix is not positive semi-definite and the optimization problem in Eq. (14) thus not convex if $\mathbf{G} = \mathbf{C}_{\text{T}} - \mathbf{C}_{\text{T}}$. However, the symmetry of the covariance difference matrix, which implies orthonormal eigenvectors $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$, can be exploited to derive a convex restriction of the objective function by letting $\left|\mathbf{C}_{\text{S}} - \mathbf{C}_{\text{T}}\right| = \mathbf{U}\,\text{diag}(|\lambda_1|, \ldots, |\lambda_p|)\mathbf{U}^{\text{T}}$ with $|\lambda_i|$ being the absolute value of the i-th eigenvalue of the covariance difference matrix [50].

First and/or second order moment-differences can also be combined to create new CTM mechanisms. For example, in [63, 64], the second-order scatter information from unlabeled target spectra was coupled with first-order moment differences to yield the following heavily-parameterized linear system:

$$\begin{bmatrix} \mathbf{X}_{\text{S}} \\ \tau_1\mathbf{G} \\ \tau_2\boldsymbol{\mu}_{\text{diff}} \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{y}_{\text{S}} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \Leftrightarrow \quad \min_{\mathbf{b}} ||\mathbf{X}_{\text{S}}\mathbf{b} - \mathbf{y}_{\text{S}}||_2^2 + \tau_1^2\,||\mathbf{G}\mathbf{b}||_2^2 + \tau_2^2\,||\boldsymbol{\mu}_{\text{diff}}\mathbf{b}||_2^2\,, \quad \begin{matrix} \mathbf{G} = \mathbf{F}\mathbf{V}^{\text{T}} \\ \boldsymbol{\mu}_{\text{diff}} = \boldsymbol{\mu}_{\text{T}}^{\text{x}} - \boldsymbol{\mu}_{\text{S}}^{\text{x}} \end{matrix} \tag{15}$$

where $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\text{T}}$ is the SVD of the mean-centered target spectra and $\mathbf{F} = \text{diag}(f_1, \ldots, f_{n_{\text{T}}})$, $f_i = \sqrt{\gamma\sigma_i^2/(\sigma_i^2 + \gamma)}$, $\gamma > 0$. As with GLS, the particular diagonal element $0 \leq f_i < \sigma_i$ (σ_i denoting the i^{th} singular value) damps the contribution of the corresponding loading vector when γ is small; otherwise if large, the contribution is neglible. In this case, the regularized least squares system is decomposed into three components (or terms): the model misfit of the labeled source data, the penalty term associated with the scatter matrix of the unlabeled target spectra, and the penalty term associated with the shift in means between the source and target spectra.

Local Mean-Centering As already stated in section 2.2, labeled samples are usually required in order to account for conditional shifts since source and target domains have different underlying labeling functions. If one has labeled target domain samples, then Local Mean-Centering (LMC) is a deceptively simple yet effective CTM technique that has been studied e.g. in [65, 30, 34]. First, one locally mean-centers the samples such that $\mathbf{X}_{\text{S}} := \mathbf{X}_{\text{S}} - \mathbf{1}\boldsymbol{\mu}_{\text{S}}^{\text{x}}$, $\mathbf{y}_{\text{S}} := \mathbf{y}_{\text{S}} - \mathbf{1}\mu_{\text{S}}^{y}$, $\mathbf{X}_{\text{T}} := \mathbf{X}_{\text{T}} - \mathbf{1}\boldsymbol{\mu}_{\text{T}}^{\text{x}}$ and $\mathbf{y}_{\text{T}} := \mathbf{y}_{\text{T}} - \mathbf{1}\mu_{\text{T}}^{y}$. In Eq. (8), we assign $\mathbf{G}$ and $\mathbf{h}$ to $\mathbf{X}_{\text{T}}$ and $\mathbf{y}_{\text{T}}$, respectively and apply an arbitrary regression method to

$$\begin{bmatrix} \mathbf{X}_{\text{S}} \\ \tau\mathbf{X}_{\text{T}} \end{bmatrix} \mathbf{b} = \begin{bmatrix} \mathbf{y}_{\text{S}} \\ \tau\mathbf{y}_{\text{T}} \end{bmatrix}. \tag{16}$$

Spectroscopically, solving Eq. (16) enlarges the pool of samples by labeled target domain samples, whereby the unmodeled variance not present in the source samples will be be incorporated into $\mathbf{b}$. Mechanistically, the initial local mean-centering step separately moves both the source and target centroids to the origin (in both the spectra X and label Y spaces). Hence, at this stage, we are guaranteed some amount of overlap in the marginal distributions of source and target domains. The subsequent multiplication by τ in Eq. (16) allows one to better control the amount of domain overlap in $P(X)$ and $P(Y)$: when $\tau > 1$ ($0 < \tau < 1$), we enlarge (shrink) the target space to better match that of the source space. However, from a Procrustes analysis point of view, LMC is incomplete as it invokes translation and scaling but not rotation in the alignment process, which leads to incomplete alignment of $P_{\text{T}}(X)$ to match $P_{\text{S}}(X)$. Nonetheless, LMC remains reasonably effective in that combining source and target samples into a single pool can (partially) achieve all three types of shifts: covariate, prior and conditional. An even more effective way to account for covariate and conditional shifts simultaneously is to combine LMC with the ideas from Eq. (14), i.e. augment the calibration set with labeled target domain samples and control for first and second order moment differences between source and target domain by means of regularization (see also section 4.1.1).

Generalized Singular Value Decomposition In GTR, When $\mathbf{G} \neq \mathbf{I}$ the solution vector of Eq. (9) cannot be expressed as a linear combination of loading vectors of the source spectra $\mathbf{X}_{\text{S}}$ alone. However, Generalized Singular Value Decomposition (GSVD) can be used to express the solution in terms of a shared basis set between $\mathbf{X}_{\text{S}}$ and $\mathbf{G}$. There are many forms that the GSVD can take, and they depend upon the dimensions of the matrices $\mathbf{X}_{\text{S}}$ and $\mathbf{G}$—see [58] for the common chemometrics case when the number of rows (samples) typically does not exceed the number of features (wavelengths). As was done in Eq. (10), we replace $\mathbf{G}$ with the new penalty matrix $\mathbf{H} = (1-\alpha)\mathbf{P}_{\text{G}} + \alpha\mathbf{P}_{\text{G}}^{\perp}$. As a consequence, we can now define the GSVD of $\mathbf{X}_{\text{S}}$ and $\mathbf{H}$:

$$\mathbf{X}_{\text{S}} = \mathbf{U}_{\text{S}}[\mathbf{0}\ \boldsymbol{\Sigma}_{\text{S}}]\widetilde{\mathbf{W}}^{-1} \ \text{ and } \ \mathbf{H} = \mathbf{U}_{\text{H}} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\text{H}} \end{bmatrix} \widetilde{\mathbf{W}}^{-1}, \ \ \widetilde{\mathbf{W}} = [\mathbf{W}_{\emptyset}\ \ \mathbf{W}] \tag{17}$$

where $\mathbf{U}_{\text{S}}$ and $\mathbf{U}_{\text{H}}$ are orthonormal matrices of size $n_{\text{S}} \times n_{\text{S}}$ and $p \times p$, respectively; $\boldsymbol{\Sigma}_{\text{S}} = \text{diag}(\sigma_{(1,1)}, \ldots, \sigma_{(1,n_{\text{S}})})$ and $\boldsymbol{\Sigma}_{\text{H}} = \text{diag}(\sigma_{(H,1)}, \ldots, \sigma_{(H,n_{\text{S}})})$ are diagonal matrices of size $n_{\text{S}} \times n_{\text{S}}$; and $\widetilde{\mathbf{W}}$ is a $p \times p$ invertible matrix (not necessarily orthogonal) such that $\mathbf{W}_{\emptyset}$ corresponds to the first $p - n_{\text{S}}$ columns of $\widetilde{\mathbf{W}}$ while $\mathbf{W}$ corresponds to the remaining columns. Given this decomposition, Eq. (9) can then be defined (after some algebra) as follows [56]:

$$\begin{aligned} \mathbf{b} &= \left(\mathbf{X}_{\text{S}}^{\text{T}}\mathbf{X}_{\text{S}} + \tau^2\mathbf{G}^{\text{T}}\mathbf{G}\right)^{-1}\mathbf{X}_{\text{S}}^{\text{T}}\mathbf{y}_{\text{S}} = \mathbf{W}\mathbf{F}\mathbf{c} = \sum_{i=1}^{n_{\text{S}}} f_i c_i \mathbf{w}_{\text{i}}; \\ \mathbf{F} &= \text{diag}(f_1, \ldots, f_{n_{\text{S}}}),\ f_i = \frac{\gamma_i^2}{\gamma_i^2 + \tau^2},\ \gamma_i = \frac{\sigma_{(S,i)}}{\sigma_{(H,i)}};\ \ c_i = \frac{\mathbf{u}_{(\text{S,i})}^{\text{T}}\mathbf{y}_{\text{S}}}{\sigma_{(S,i)}} \end{aligned} \tag{18}$$

Historically, the values $\gamma_i = \sigma_{(S,i)}/\sigma_{(H,i)}$ are referred to as the generalized singular values such that $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_{n_{\text{S}}}$ (which is the reverse of the ordering normally associated with the SVD). From the GSVD perspective, each shared basis vector $\mathbf{w}_i$ has a fuzzy membership: the spectral behavior of $\mathbf{w}_i$ is either predominantly associated or characterized by $\mathbf{X}_{\text{S}}$, $\mathbf{G}$ or both. According to [66], if $\gamma_i \gg 1$ (or $\sigma_{(S,i)} \gg \sigma_{(H,i)}$), then the basis vector behavior of $\mathbf{w}_i$ is dominated by $\mathbf{X}_{\text{S}}$; otherwise, if $\gamma_i \ll 1$ (or $\sigma_{(S,i)} \ll \sigma_{(H,i)}$), then $\mathbf{G}$ dominates $\mathbf{w}_i$. When $\gamma_i \approx 1$, then $\mathbf{w}_i$ is equally characterized by both $\mathbf{X}_{\text{S}}$ and $\mathbf{G}$. Moreover, each diagonal element $0 < f_i \leq 1$ in $\mathbf{F}$ damps the contribution of $\mathbf{w}_i$ when the penalty parameter τ is large. (The basis vectors $\mathbf{w}_i$ associated with small generalized eigenvectors γ_i will be disproportionally damped or filtered relative to basis vectors associated with large generalized eigenvectors.)

With respect to CTM applications, the GSVD has been tangentially used. Recall that in Eq. (15), the matrix $\mathbf{G}$ was set to a second-moment scatter matrix associated with the target spectra $\mathbf{X}_{\text{T}}$ [63]. Although there was no mention of GSVD in [63] (nor in [64], the paper which originally introduced the penalized least squares framework of Eq.(15) and was subsequently extended by [63]), their work has a direct relationship with GSVD: the regression vector $\mathbf{b}$ is a linear combination of loading vectors $\mathbf{w}_i$ where each loading vector is dominated by either the source spectra (when $\gamma_i \gg 1$), the target spectra (when $\gamma_i \ll 1$) or both (when $\gamma_i \approx 1$). To emphasize, when $\mathbf{G}$ is not associated with domain difference spectra but instead associated with target spectra, then CTM via GSVD takes on characteristics more associated with data fusion.

3.2.2 Domain-invariant partial least squares regression (di-PLS)

The development of di-PLS was strongly inspired by the subspace-based DA methods introduced in Section 3.1.2. The core idea behind di-PLS is to identify a subspace that is predictive w.r.t. the response in the source domain and where the distributional difference in terms of first and second order moments (i.e. mean and co-variance) between source and target domain spectra is small. In fact, di-PLS can be viewed in terms of both, GTR and GE frameworks depending on the objective function (NIPALS or covariance maximization) one resorts to for deriving the (PLS) weight vector:

$$\min_{\mathbf{w}} \|\mathbf{X} - \mathbf{y}\mathbf{w}^{\mathrm{T}}\|_F^2 + \tau \mathbf{w}^{\mathrm{T}}\mathbf{G}\mathbf{w} \quad \Rightarrow \; \mathbf{w} = \left(\mathbf{I} + \frac{\tau}{\mathbf{y}^{\mathrm{T}}\mathbf{y}}\mathbf{G}\right)^{-1} \frac{\mathbf{X}^{\mathrm{T}}\mathbf{y}}{\mathbf{y}^{\mathrm{T}}\mathbf{y}} \qquad \text{(GTR)} \qquad (19)$$

$$\max_{\mathbf{w}} \; \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} \quad \text{s.t.} \quad \mathbf{w}^{\mathrm{T}}(\mathbf{G} + \delta\mathbf{I})\mathbf{w} = 1 \quad \Rightarrow \; \mathbf{X}^{\mathrm{T}}\mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \lambda(\mathbf{G} + \delta\mathbf{I})\mathbf{w} \qquad \text{(GE)} \qquad (20)$$

Similar as in Eq. (14), $\mathbf{G} = \left|\mathbf{C}_{\mathrm{S}} - \mathbf{C}_{\mathrm{T}}\right|$ encodes the co-variance difference between the domains. Expanding the regularization term in Eq. (19) accordingly yields $\mathbf{w}^{\mathrm{T}}\mathbf{G}\mathbf{w} \geq \left|\frac{1}{n_{\mathrm{S}}}\mathbf{w}^{\mathrm{T}}\mathbf{X}_{\mathrm{S}}^{\mathrm{T}}\mathbf{X}_{\mathrm{S}}\mathbf{w} - \frac{1}{n_{\mathrm{T}}}\mathbf{w}^{\mathrm{T}}\mathbf{X}_{\mathrm{T}}^{\mathrm{T}}\mathbf{X}_{\mathrm{T}}\mathbf{w}\right|$ which, in case of locally mean centered matrices $\mathbf{X}_{\mathrm{S}}$ and $\mathbf{X}_{\mathrm{T}}$, is equivalent to (an upper bound on) the absolute difference between the variance of source and target domain in the direction $\mathbf{w}$ [49]. The GE type formulation in Eq. (20), on the other hand, reveals that the (domain-invariant) weight vector $\mathbf{w}$ is the generalized eigenvector of $\mathbf{X}^{\mathrm{T}}\mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{X}$ with respect to $(\mathbf{G} + \delta\mathbf{I})$ (with δ being a regularization parameter) that is associated with the largest eigenvalue. Huang *et al.* recently proposed to replace $\mathbf{G}$ in Eq. (20) with a dedicated Laplacian matrix to account for the distributional differences in a non-parameteric way (i.e. without the need to estimate source and target covariance matrices), which can be beneficial if data is scarce and/or the distributional differences are more complicated [67]. Python code for di-PLS is publicly available under `https://github.com/B-Analytics/di-PLS`.

4 Applications

In this section, we apply CTM and DA techniques to two spectral datasets from the public-domain involving calibration transfer between similar instruments and a simulated data set involving a change in sample matrix. The former involve (mostly) covariate shifts, while the latter deals with conditional shifts.

4.1 Common CTM/DA Approaches Applied to CORN and SOY

Corn Dataset The corn instrument data set consists of eighty NIR spectra measured from three instruments labeled m5, mp5, and mp6 across 700 wavelengths between the spectral region of 1100 to 2498 nm at 2 nm intervals [2]. The source and target samples are drawn from the m5 and mp6 instruments, respectively; the most dissimilar instruments. For every sample, four response variables were measured: moisture, oil, protein, and starch. We used protein as the response variable.

There is no default or designated split of the data into training and test sets. As a result, we created 100 random sample splits of the data. Each data split is disjoint and partitioned in the following manner.

- Set $\mathcal{C}$: Forty source samples were used as the *calibration* or training set.
- Set $\mathcal{V}$: A distinct set of thirty-five target samples were set aside and used as the *validation* or test set.
- Set $\mathcal{R}$: The *remaining* five samples were used (or not used) as matched samples, labeled samples or unlabeled samples—depending upon the CTM approach.

The various CTM approaches will now be briefly discussed.

- **NONE**: No calibration updating was performed. A PLS model was build on $\mathcal{C}$ and predictions were made on $\mathcal{V}$. The five samples in $\mathcal{R}$ were not used.
- **Piecewise Direct Standardization (PDS)**: For PDS, the transfer matrix $\mathbf{F}$ mapping spectra from mp6 to m5 was constructed using the matched set of five samples from both m5 and mp6 in $\mathcal{R}$.
- **Generalized Eigenproblem (GE)**: The source spectra from $\mathcal{C}$ and target spectra (mp6) from $\mathcal{R}$ were used to construct the Laplacian matrix based upon Eq. (4).
- **Generalized Eigenproblem in Transudctive mode (GE-T)**: The source spectra from $\mathcal{C}$ and target spectra (mp6) from both $\mathcal{R}$ and $\mathcal{V}$ were used to construct the Laplacian matrix based upon Eq. (4). For GE, it is easy to accommodate unlabeled target samples.

[2] http://www.eigenvector.com/data/Corn, accessed Jan. 2021

- **Matched Samples (MS)**: Here, the difference spectra was constructed from the matched set of five samples from both m5 and mp6 in $\mathcal{R}$. This difference matrix was used as the GTR matrix $\mathbf{G}$ in Eq. (9).
- **Local Mean centering (LMC)**: Here, the labeled target samples (both spectra and corresponding reference values) from $\mathcal{R}$ were used as $\mathbf{G}$ and $\mathbf{h}$, respectively, in Eq. (16).

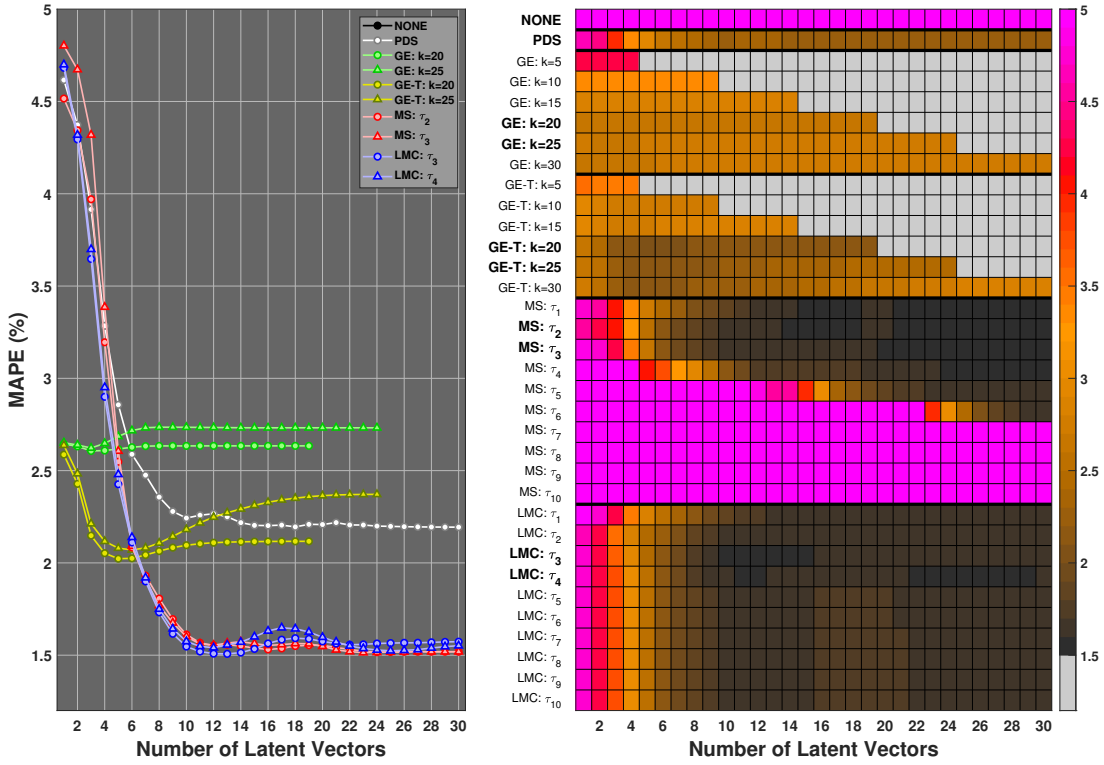

Figure 5: Performance of different CTM approaches on the corn benchmark dataset. The left subplot shows MAPE (%) on the validation set as a function of the number of PLS latent vectors. The right subplot shows the same information as a heatmap display. For purposes of de-cluttering the MAPE curves in the left subplot, only a select set of tuning parameters were shown for GE, GE-T, MS and LMC. In the right subplot, each column indicates the number of latent vectors and each row corresponds to a CTM approach. Moreover, *all* tuning parameters are shown in each row of the heatmap (and the corresponding tuning parameters shown as MAPE curves are highlighted in bold)

.

The Mean Absolute Percentage Error (MAPE, %) was used as the figure of merit:

$$\text{MAPE} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left| \frac{y_i - \hat{y}_i}{y_i} \right| 100\% \tag{21}$$

Across the 100 data splits, the average MAPE on the validation set is reported in Figure 5 as a function of the number of latent PLS vectors. In the right subplot, the same information is displayed as a heatmap. (All MAPE values or cells above 5.2 are magenta in color.) Each column indicates the number of latent vectors and each row corresponds to a CTM approach. For purposes of de-cluttering the MAPE curves in the left subplot, only a select set of tuning parameters are shown for GE, GE-T, MS and LMC. On the other hand, all tuning parameters are shown in each row of the heatmap (and the corresponding tuning parameters shown as MAPE curves are highlighted in bold). For PDS, the best performace was achieved with a window width of 1. Other window widths were employed but are not shown since they yielded inferior results. For both GE and GE-T, we examined the effect of keeping the first k (and dominant) eigenvectors of $\mathbf{W}$ such that $k \in \{5, 10, 15, 20, 25, 30\}$. For MS and LMC, the τ values (the two-norm penalty associated with GTR) were used such that $\tau_1 < \tau_2 < \cdots < \tau_{10}$ and are exponentially decaying to zero.

There are a couple of trends to note in Figure 5. First, having access either to labeled target samples (LMC) or matched samples (MS) clearly results in superior performance. However, LMC performance can be deemed superior since performance (as displayed in the heatmap) is relatively insensitive to the two-norm penalty parameter τ. If one does not have access to labeled target samples, then having matched samples to work with is the next best alternative. The

generalized eigenproblem approach varies in performance—depending upon whether the unlabeled spectra in $\mathcal{R}$ alone (GE) or whether an unlabeled set of spectra is used as well (transductive mode, or GE-T) to construct the Laplacian matrix in Eq. (4). Having access to more unlabeled spectra is always preferable to less, and here, GE-T can outperform PDS provided one uses a large enough set of eigenvectors. Incorporating unlabeled spectra from the test set into the modeling updating procedure is not a far-fetched idea in chemometrics. For example, unlabeled spectra acquired from an instrument in the field—if the instrument is tethered to a phone such that spectral data can be quickly uploaded to computational cloud services—can make transductive inference more readily accessible.

The superior performance of LMC is likely due to it being the only CTM method to have access to labeled target samples. That is, in the prediction of the new sample, the target means are used: $y_{\text{new}} = (\mathbf{x}_{\text{new}} - \boldsymbol{\mu}_{\text{T}}^{\text{x}})\mathbf{b} + \mu_{\text{T}}^{\text{y}}$. However, GTR based on matched samples perform equally well or even better (depending on the choice of the regularization parameter), which indicates that the instrument differences are (mostly) characterized by covariate shift. Consequently, reference measurements from the target domain samples are not necessarily required for model maintenance. If one only has access to matched samples, then model updating via MS has a performance edge over PDS, especially if one chooses a large two-norm penalty parameter (a large two-norm parameter was also observed to be beneficial in the MS scenario described in [38]). But the reality is that having access to labeled target samples and/or matched samples is a calibration luxury in many practical applications and use-case scenarios. In the absence of such samples, then meaningfully using unlabeled target samples is still relatively unexplored territory but warrants further investigation.

4.1.1 Covariate vs conditional shifts

We will proceed by showcasing the application of DA on a simulated spectral dataset, where the goal is to adapt a multivariate calibration model from a source to a target domain (Figure 6). The source domain consists of two Gaussians centered around $\mu = 50$ (analyte) and $\mu = 65$ (interferrent 1) added together in different proportions (including some random noise), while in the target domain an additional interferrent centered around $\mu = 35$ (interferrent 2) has been added in different amounts. The distribution of the analyte concentration (i.e. $P(Y)$) remains unchanged between the domains, i.e. there is no prior shift. Obviously, it holds that $P_{\text{S}}(X) \neq P_{\text{T}}(X)$, i.e. there must be a covariate shift between the domains since interferent 2 changes the convariance structure of the data. In addition, the difference between the domains also involves a conditional shift, i.e. it holds that $P_S(Y|X) \neq P_T(Y|X)$, because the second interferrent overlaps significantly with the analyte and thus changes the correlation between predictors and response - at least around the "absorbance" peak of the interferrent. As we will see later, the severity of the conditional shift is directly associated with the amount of this overlap. We will start by correcting for the covariate shift, which requires (in addition to the source domain data including the analyte concentrations) only a (representative) set of spectra from the target domain without the corresponding reference measurements. This is the typical setting in unsupervised domain adaptation which can be accomplished either by the generalized Eigendecomposition type methods from section 3.1.2 (e.g. TCA), GTR approaches (e.g. Eq. (14)) or di-PLS (Figure 6C). All these methods involve a tuning parameter that controls how much emphasis should be given to correct for the covariate shift. This however is a non-trivial task when there are no reference values in the target domain to validate the model. An approach that has been applied with success in the past is to plot the distributional difference between the domains (in the LV space) against increasing values of the regularization parameter and choose the value where this curve is closest to the origin (Figure 6D). This can be done globally or for each LV individually [40]. A simpler (yet less flexible) alternative is to optimize the distributional difference between $\mathbf{y}_S$ and $\hat{\mathbf{y}}_T$, i.e. to choose the regularization parameter such that the distribution of the predictions on the (unlabeled) target domain samples better matches the distribution of the reference values in the source domain. In parallel, either the weight, the loadings or the regression coefficients should be inspected. If too much emphasis is placed on aligning the distributions, the risk of aligning the noise rather than the (predictive) information increases which manifests itself in these quantities (Figure 6E). At the same time it is also advisable to check how regularization affects the source error as there is a trade-off between fitting the response (in the source domain) and aligning the distributions. Unfortunately, di-PLS tends to shrink the variance of the predictors (i.e. the scores) and domain regularization thus often leads to lower errors in the source domain compared to standard PLS when the number of LVs is small [68].

An appealing feature of the GTR type methods and di-PLS is that the models can accommodate (a small fraction of) labeled target domain samples along with (a larger fraction of) unlabeled target domain samples in order to correct for both covariate and conditional shifts. Figure 6F shows the RMSEP of di-PLS models in the target domain in dependence of the number of reference measurements available in the target domain and the distance Δ between the target domain-specific interferrent's and the analyte's peak (indicated in Figure 6B). If this distance is large, the difference between the domains involves predominantly covariate shift and thus the RMSEP does not improve significantly when including label information for (some of the) target domain spectra. However, for small Δ's the conditional shift becomes more severe and the RMSEP can thus only be improved by including labeled target domain samples.

Although quite simplistic, the above example is useful to emphasize the (potential) benefit (i.e. model adaptation with unlabeled data), challenges (how to select and optimize a model when labels/reference values are scarce or not available?) and open questions (when is DA feasible?) with domain adaptation.

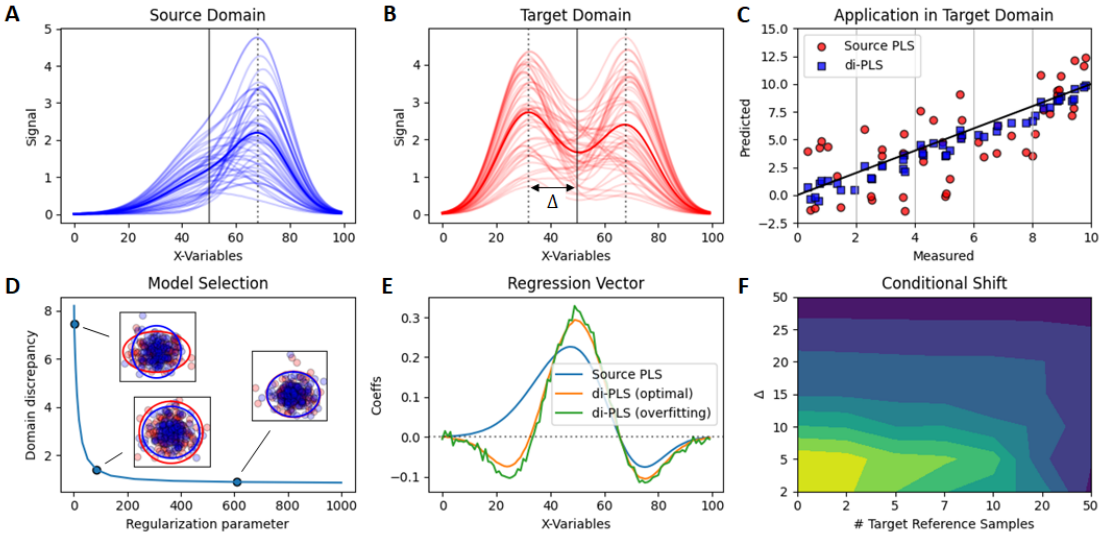

Figure 6: Application of di-PLS for domain adaptation. Simulated source (A) and target (B) domain datasets. Solid and dotted lines indicate the peaks, where analyte and interferrents have maximal signals. Δ indicates the difference between analyte and target domain-specific interferrent peak. (C) Measured vs. predicted analyte concentration in the target domain from a PLS model fitted to the source domain data only (red) and a di-PLS model fitted to labeled source and unlabeled target domain data. (D) Domain discrepancy vs. amount of regularization. The inset plots show projections of source (blue) and target (red) domain data together with the 95%-CI along the first 2 LVs of a di-PLS model at the three points indicated in the figure. (E) Regression coefficients of the source PLS model and di-PLS models at optimal regularization and when regularization is too strong (middle and rightmost points in D). (F) Dependence of root mean squared error of prediction (RMSEP) on the number of reference measurements available for the target domain samples and distance Δ between analyte and target domain-specific interferrent peak. Blue and yellow regions indicate low and high RMSEPs, respectively.

5 Perspectives

The previous section has shown that DA – if applied with the concepts from section 2 in mind – can be highly useful when it comes to adapting multivariate calibrations between related domains. DA and recent methods developed in the field of chemometrics (e.g. di-PLS) hold promise to help solving such types of real-world problems in analytical chemistry. Moreover, those DA methods that do not require, or at least reduce the need for, costly and time consuming laboratory analyses of reference samples are of particular interest. However, there is an inherent tension when using unsupervised DA (no reference measurements available in the target domain). Theory shows that even for simple covariate shifts, unsupervised DA of the source model to the target domain can (in principle) fail [7]. On the other hand, unsupervised DA models might generalize sufficiently well across two (or more) domains despite conditional shift. In the example shown in Figure 6 the ability of the model to generalize to the target domain can be attributed to the presence of enough selective information about the analyte in both domains. In short, knowledge about the distributional characteristics of the domain differences (i.e. covariate/conditional/prior shift) is not sufficient. Knowledge about information associated with the analyte and interferents is also imperative to develop a model adaptation strategy. Regardless of the algorithmic approaches at one's disposal, a more fundamental question has to be asked: Is DA even feasible at all, and if so, are reference measurements required in the target domain? Unfortunately, measures or figures of merit that assess the feasibility of DA in a given practical situation are as yet largely missing in the chemometric literature. Thus, future work should focus on the development of such measures.

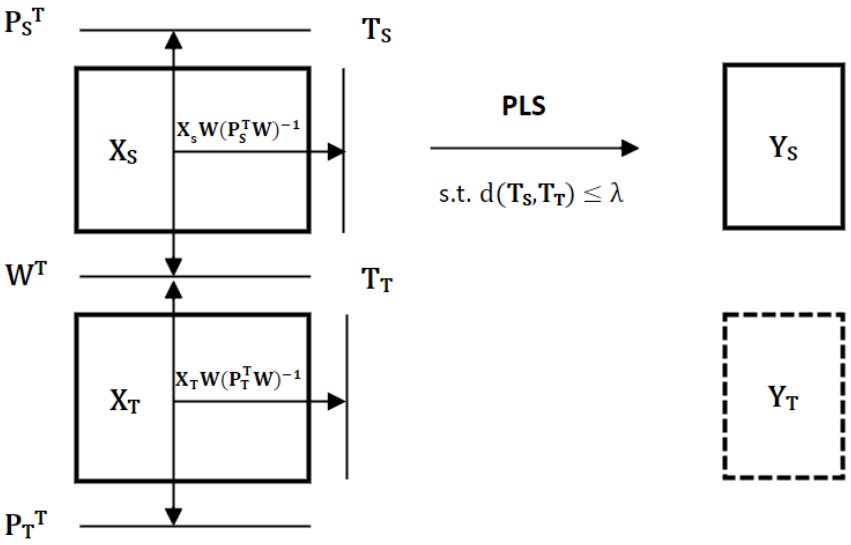

Figure 7: Data integration interpretation of di-PLS. Schematic drawing of di-PLS regression of $\mathbf{Y}_S$ on $\mathbf{X}_S$ such that the distributional difference $d(\mathbf{T}_S, \mathbf{T}_T)$ between the projections of $\mathbf{X}_S$ (source domain) and $\mathbf{X}_T$ (target domain) are not larger then some pre-defined λ. Note that source and target domain matrices have domain-specific scores and loadings matrices $\mathbf{P}$ and $\mathbf{T}$ but share a single weight matrix $\mathbf{W}$. When $\mathbf{Y}_T$ is not available, di-PLS addresses the unsupervised domain adaptation problem.

5.1 Data Integration and Fusion

Similar to the DA methods discussed in Section 3, the aim of data integration/fusion methods is to identify common, latent phenomena in related data sets [69]. In fact, data integration methods have been successfully employed in the past for calibration transfer, i.e. to transfer calibrations between similar spectrometers [70] and domain adaptation [63]. On the other hand, some of the DA methods discussed in section 3 can be regarded as *striving* for data integration. In Section 3.2.1, GTR, when viewed through the prism of GSVD, creates a common basis set that spans both the source and target domains. With respect to di-PLS, Figure 7 shows a simplified LV model scheme where the goal is to regress the response to the predictors in a source domain under the constraint that the distributional difference between the domain-specific latent representations (i.e. the scores) is small.

The differences between source and target samples follow a continuum between dissimilarity and similarity. If the source and target samples are highly similar (i.e., are drawn from the same underlying distribution and share the same sources of variation), then a source model will likely perform well in the target domain. If the target samples are radically dissimilar from the source samples, then one can ignore the source samples and build a calibration model solely on the target samples (provided there are enough samples). Many settings in chemometrics are intermediate between these two extremes: although we expect the target samples to be dissimilar (but related) to the source samples, the source samples should still provide leverage such that improved prediction can be achieved for target samples. However, how do we quantify how much common information is present across domains in this continuum? And is this common information predictive with respect to the response? Recent work has made a preliminary attempt at this issue for domain adaptation situations [71]. Regarding the feasibility of unsupervised DA, another relevant question is the following: Under which (practical) conditions is a "common" representation of the domains (in a data fusion sense) invariant with respect to their distributional properties (e.g. mean, co-variance etc.) or vice versa. After all, generalization across the source and target domain statistically implies that the samples were at least sampled from a common underlying distribution. Moreover, one further assumes that the common (latent) information is also predictive with respect to the response. In our opinion, future work should focus on sheding light on the relationship between the concepts of common and domain-invariant LVs.

5.2 Transfer Learning

As already stated in the introduction, most lines of work in chemometrics have thus far addressed domain adaptation rather than transfer learning. The main difference between DA and TL is that in the latter, data from related domains is leveraged to learn a new task instead of learning the same task in a new domain. In some computer vision

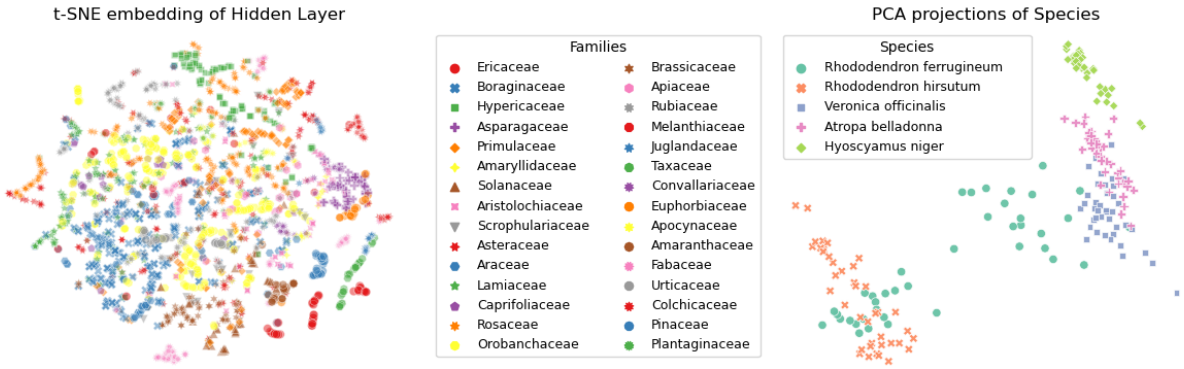

Figure 8: Deep transfer learning on ATR-FTIR spectra of dried plant leafs. Left plot: t-SNE embeddings of the activations from the last hidden layer of a fully connected deep neural network with 4 hidden layers trained on a dataset containing spectra from 30 distinct plant families (unpublished work). Right plot: Projections of samples from 5 species (not included during training) on the first two principle components fitted to the hidden activations from the training data.

problems, TL is possible because some high-level features (e.g. wheels) extracted while learning a task in one domain (e.g. discrimination between trucks and cars) can help learning a semantically related task in another domain (e.g. discrimination between motorbikes and bicycles). It is reasonable to assume that such high-level features can also be derived by means of deep learning approaches, e.g., from NIR spectra reflecting general classes of vibrational modes that characterize similar kinds of molecules. However, in contrast to computer vision problems, where individual objects can usually be spatially isolated from each other and the background, in spectroscopy, the signals from individual molecules often overlap heavily with the background signal from the sample matrix. In addition, TL in computer vision and related domains is employed predominantly for classification tasks, while most applications in chemometrics involve regression problems and deal with quantitative analysis of molecules. While DL has been successfully employed for multivariate calibration and domain adaptation in chemometrics with the type of models usually employed in computer vision (e.g. convolutional neural networks) [72], true transfer learning with spectroscopic data will require large compilations of datasets to learn domain-specific feature representations for data from a particular analytical platform (e.g. NIR or MIR spectroscopy). Figure 8 exemplifies this idea of transfer learning using a dataset of ATR-FTIR spectra of dried plant leaf surfaces from 30 distinct plant families (unpublished work) [73, 74]. The left plot shows non-linear t-SNE [75] embeddings (i.e. projections) of the activations (i.e. the scores) of the last hidden layer from a fully connected deep neuronal network fitted to a subset of this dataset. As can be seen from this plot, most of the data points belonging to the same plant family are close to each other indicating that the hidden layers encode discriminating features between the classes. The right plot shows projections of some plant species (that were not included during training of the DNN) onto the first two principle components of a PCA model fitted to the same hidden activations. This is to show that the same features that are useful to discriminate between plant families (i.e. the source task) turn out to be useful (to some extent) for the discrimination between different genera from different plant families (e.g. *H. niger* and *V. officinalis*), different genera from the same family (*A. belladonna* and *H. niger*, both Solanaceae) and between different species from the same genus (*R. ferrugineum* and *R. hirsutum*). Just like a botanical novice, the initial DNN first learns the features that characterize different plant families. Some of this knowledge might help right away to distinguish plants from different genera of the same family. However, further training might be necessary in order to tell closely related species apart safely.

6 Conclusion

The primary aim of this contribution was to give the average reader with a background in chemometrics a "gentle primer" on the subject of transfer learning and domain adaptation and to offer an alternative perspective on some of the methods that chemometricians have developed over the past decades. We hope that our contribution will help practitioners to tackle real-world problems in analytical chemistry involving calibration maintenance and transfer more efficiently and foster new developments in chemometrics and analytical chemistry.

Acknowledgements

The first author acknowledges funding from the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK), the Federal Ministry for Digital and Economic Affairs (BMDW), and the Province of Upper Austria in the frame of the COMET - Competence Centers for Excellent Technologies programme managed by the Austrian Research Promotion Agency FFG, the COMET Center CHASE and the FFG project Interpretable and Interactive Transfer Learning in Process Analytical Technology (Grant No. 883856). We further thank Dr. Florian Sobieczky for fruitful discussions and the first reviewer for helping to improve the manuscript.

References

[1] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[3] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016.

[4] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[6] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[7] Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

[8] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *ICML*, pages 2839–2848, 2016.

[9] B. G. OSBORNE and T. FEARN. Collaborative evaluation of universal calibrations for the measurement of protein and moisture in flour by near infrared reflectance. *International Journal of Food Science & Technology*, 18(4):453–460, 1983.

[10] J. S. Shenk, M. O. Westerhaus, and W. C. Templeton Jr. Calibration transfer between near infrared reflectance spectrophotometers1. *Crop Science*, 25(1):cropsci1985.0011183X002500010038x, 1985.

[11] MO Westerhaus. Improving repeatability of nir calibrations across instruments. In R. Biston and Eds. N. Bartiaux-Thill, editors, *Proceedings of the Third International Near Infrared Spectroscopy Conference*, page p. 671, Gembloux, Belgium, 1991. Agriculture Research Centre Publishing.

[12] Yongdong. Wang, David J. Veltkamp, and Bruce R. Kowalski. Multivariate instrument standardization. *Analytical Chemistry*, 63(23):2750–2756, 1991.

[13] Yongdong. Wang and Bruce R. Kowalski. Temperature-compensating calibration transfer for near-infrared filter instruments. *Analytical Chemistry*, 65(9):1301–1303, 1993.

[14] Onno E. de Noord. Multivariate calibration standardization. *Chemometrics and Intelligent Laboratory Systems*, 25(2):85–97, 1994.

[15] Onno E. de Noord. The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems*, 23(1):65–70, 1994. Proceedings of the 3rd Scandinavian Symposium on Chemometrics (SSC3).

[16] Edward V. Thomas. Incorporating auxiliary predictor variation in principal component regression models. *Journal of Chemometrics*, 9(6):471–481, 1995.

[17] E. Bouveresse and D.L. Massart. Improvement of the piecewise direct standardisation procedure for the transfer of nir spectra for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 32(2):201–213, 1996.

[18] Svante Wold, Henrik Antti, Fredrik Lindgren, and Jerker Öhman. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44(1):175–185, 1998.

[19] Jonas Sjöblom, Olof Svensson, Mats Josefson, Hans Kullberg, and Svante Wold. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44(1):229–244, 1998.

[20] H. Swierenga, W. G. Haanstra, A. P. De Weijer, and L. M. C. Buydens. Comparison of two different approaches toward model transferability in nir spectroscopy. *Appl. Spectrosc.*, 52(1):7–16, Jan 1998.

[21] Chad E. Anderson and John H. Kalivas. Fundamentals of calibration transfer through procrustes analysis. *Applied Spectroscopy*, 53(10):1268–1276, 1999.

[22] David M. Haaland and David K. Melgaard. New prediction-augmented classical least-squares (pacls) methods: Application to unmodeled interferents. *Applied Spectroscopy*, 54(9):1303–1312, 2000.

[23] Tom Fearn. Standardisation and calibration transfer for near infrared instruments: A review. *Journal of Near Infrared Spectroscopy*, 9(4):229–244, 2001.

[24] David M Haaland and David K Melgaard. New augmented classical least squares methods for improved quantitative spectral analyses. *Vibrational Spectroscopy*, 29(1):171–175, 2002. A Collection of papers presented at the 1st International Conference on Advanced Vibrational Spectroscopy, Turku, Finland, August 19-24, 2001.

[25] Christine M. Wehlburg, David M. Haaland, David K. Melgaard, and Laura E. Martin. New hybrid algorithm for maintaining multivariate quantitative calibrations of a near-infrared spectrometer. *Appl. Spectrosc.*, 56(5):605–614, May 2002.

[26] Christopher D. Brown. Discordance between net analyte signal theory and practical multivariate calibration. *Analytical Chemistry*, 76(15):4364–4373, 2004. PMID: 15283574.

[27] Harald Martens, Martin Høy, Barry M. Wise, Rasmus Bro, and Per B. Brockhoff. Pre-whitening of data by covariance-weighted pre-processing. *Journal of Chemometrics*, 17(3):153–165, 2003.

[28] Yusuf Sulub and Gary W. Small. Spectral simulation methodology for calibration transfer of near-infrared spectra. *Appl. Spectrosc.*, 61(4):406–413, Apr 2007.

[29] Benoit Igne and Charles R. Hurburgh. Standardisation of near infrared spectrometers: Evaluation of some common techniques for intra- and inter-brand calibration transfer. *J. Near Infrared Spectrosc.*, 16(6):539–550, Dec 2007.

[30] John H. Kalivas, Gabriel G. Siano, Erik Andries, and Hector C. Goicoechea. Calibration maintenance and transfer using tikhonov regularization approaches. *Applied Spectroscopy*, 63(7):800–809, 2009. PMID: 19589218.

[31] Paman Gujral, Michael Amrhein, Barry M. Wise, and Dominique Bonvin. Framework for explicit drift correction in multivariate calibration models. *Journal of Chemometrics*, 24(7-8):534–543, 2010.

[32] M. Ross Kunz, John H. Kalivas, and Erik Andries. Model updating for spectral calibration maintenance and transfer using 1-norm variants of tikhonov regularization. *Analytical Chemistry*, 82(9):3642–3649, 2010.

[33] Erik Andries and John H. Kalivas. Interrelationships between generalized tikhonov regularization, generalized net analyte signal, and generalized least squares for desensitizing a multivariate calibration to interferences. *Journal of Chemometrics*, 27(5):126–140, 2013.

[34] Erik Andries. Penalized eigendecompositions: motivations from domain adaptation for calibration transfer. *Journal of Chemometrics*, 31(4):e2818, 2017. e2818 cem.2818.

[35] Jr. Jerome J. Workman. A review of calibration transfer practices and instrument differences in spectroscopy. *Applied Spectroscopy*, 72(3):340–365, 2018. PMID: 28929781.

[36] Ziyi. Wang, Thomas. Dean, and Bruce R. Kowalski. Additive background correction in multivariate instrument standardization. *Analytical Chemistry*, 67(14):2379–2385, 1995.

[37] Yongdong Wang and Bruce R. Kowalski. Calibration transfer and measurement stability of near-infrared spectrometers. *Appl. Spectrosc.*, 46(5):764–771, May 1992.

[38] Ramin Nikzad-Langerodi and Florian Sobieczky. Graph-based calibration transfer. *Journal of Chemometrics*, 35(4):e3319, 2021. e3319 cem.3319.

[39] Puneet Mishra, Ramin Nikzad-Langerodi, Federico Marini, Jean Michel Roger, Alessandra Biancolillo, Douglas N. Rutledge, and Santosh Lohumi. Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? the answer is not always. *TrAC Trends in Analytical Chemistry*, 143:116331, 2021.

[40] Ramin Nikzad-Langerodi, Werner Zellinger, Edwin Lughofer, and Susanne Saminger-Platz. Domain-invariant partial-least-squares regression. *Analytical Chemistry*, 90(11):6693–6701, 2018. PMID: 29722978.

[41] Puneet Mishra and Ramin Nikzad-Langerodi. Partial least square regression versus domain invariant partial least square regression with application to near-infrared spectroscopy of fresh fruit. *Infrared Physics & Technology*, 111:103547, 2020.

[42] Puneet Mishra and Ramin Nikzad-Langerodi. A brief note on application of domain-invariant pls for adapting near-infrared spectroscopy calibrations between different physical forms of samples. *Talanta*, 232:122461, 2021.

[43] Ramin Nikzad-Langerodi, Edwin Lughofer, Carlos Cernuda, Thomas Reischer, Wolfgang Kantner, Marcin Pawliczek, and Markus Brandstetter. Calibration model maintenance in melamine resin production: Integrating drift detection, smart sample selection and model adaptation. *Analytica Chimica Acta*, 1013:1–12, 2018.

[44] Thomas Boucher, M Darby Dyar, and Sridhar Mahadevan. Proximal methods for calibration transfer. *Journal of Chemometrics*, 31(4):e2877, 2017.

[45] Yehuda Koren and Liran Carmel. Robust linear dimensionality reduction. *IEEE transactions on visualization and computer graphics*, 10(4):459–470, 2004.

[46] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

[47] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

[48] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

[49] Ramin Nikzad-Langerodi, Werner Zellinger, Susanne Saminger-Platz, and Bernhard Moser. Domain-invariant regression under beer-lambert's law. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 581–586, 2019.

[50] Ramin Nikzad-Langerodi, Werner Zellinger, Susanne Saminger-Platz, and Bernhard A Moser. Domain adaptation for regression under beer–lambert's law. *Knowledge-Based Systems*, 210:106447, 2020.

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

[52] Jean-Michel Roger and Jean-Claude Boulet. A review of orthogonal projections for calibration. *Journal of Chemometrics*, 32(9):e3045, 2018. e3045 cem.3045.

[53] Jean-Claude Boulet and Jean-Michel Roger. Pretreatments by means of orthogonal projections. *Chemometrics and Intelligent Laboratory Systems*, 117:61–69, 2012. Special Issue Section: Selected Papers from the 1st African-European Conference on Chemometrics, Rabat, Morocco, September 2010 Special Issue Section: Preprocessing methods Special Issue Section: Spectroscopic imaging.

[54] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995.

[55] Åke Björck. *Numerical methods for least squares problems*. SIAM, 1996.

[56] Per Christian Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.

[57] Richard C Aster, Brian Borchers, and Clifford H Thurber. *Parameter estimation and inverse problems*. Elsevier, 2018.

[58] Timothy W. Randolph, Jaroslaw Harezlak, and Ziding Feng. Structured penalties for functional linear models-partially empirical eigenvectors for regression. *Electronic journal of statistics*, 6:323–353, Jan 2012. 22639702[pmid].

[59] X Capron, B Walczak, OE De Noord, and DL Massart. Selection and weighting of samples in multivariate regression model updating. *Chemometrics and intelligent laboratory systems*, 76(2):205–214, 2005.

[60] David K. Melgaard, David M. Haaland, and Christine M. Wehlburg. Concentration residual augmented classical least squares (cracls): A multivariate calibration method with advantages over partial least squares. *Applied Spectroscopy*, 56(5):615–624, 2002.

[61] Christine M Wehlburg, David M Haaland, and David K Melgaard. New hybrid algorithm for transferring multivariate quantitative calibrations of intra-vendor near-infrared spectrometers. *Applied spectroscopy*, 56(7):877–886, 2002.

[62] Erik Andries, John H Kalivas, and Anit Gurung. Sample and feature augmentation strategies for calibration updating. *Journal of Chemometrics*, 33(1):e3080, 2019.

[63] Jacob Søgaard Larsen, Line Clemmensen, Anders Stockmarr, Thomas Skov, Anders Larsen, and Bjarne Kjær Ersbøll. Semi-supervised covariate shift modelling of spectroscopic data. *Journal of Chemometrics*, 34(3):e3204, 2020. e3204 cem.3204.

[64] Kenneth Joseph Ryan and Mark Vere Culp. On semi-supervised linear regression in covariate shift problems. *The Journal of Machine Learning Research*, 16(1):3183–3217, 2015.

[65] Chris L. Stork and Bruce R. Kowalski. Weighting schemes for updating regression models—a theoretical approach. *Chemometrics and Intelligent Laboratory Systems*, 48(2):151–166, 1999.

[66] Orly Alter, Patrick O Brown, and David Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences*, 100(6):3351–3356, 2003.

[67] Guangzao Huang, Xiaojing Chen, Limin Li, Xi Chen, Leiming Yuan, and Wen Shi. Domain adaptive partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 201:103986, 2020.

[68] B. M. Wise and N. L. Ricker. Identification of finite impulse response models with continuum regression. *Journal of Chemometrics*, 7(1):1–14, 1993.

[69] Age K Smilde, Ingrid Måge, Tormod Naes, Thomas Hankemeier, Mirjam Anne Lips, Henk AL Kiers, Ervim Acar, and Rasmus Bro. Common and distinct components in data fusion. *Journal of Chemometrics*, 31(7):e2900, 2017.

[70] Tomas Skotare, David Nilsson, Shaojun Xiong, Paul Geladi, and Johan Trygg. Joint and unique multiblock analysis for integration and calibration transfer of nir instruments. *Analytical chemistry*, 91(5):3516–3524, 2019.

[71] A. Gurunf and John H. Kalivas. Model selection challenges with application to multivariate calibration updating methods. *Journal of Chemometrics.*, 34:e3245, 2020.

[72] Puneet Mishra and Dário Passos. Realizing transfer learning for updating deep learning models of spectral data to be used in new scenarios. *Chemometrics and Intelligent Laboratory Systems*, 212:104283, 2021.

[73] Ramin Nikzad-Langerodi, Katharina Arth, Valerie Klatte-Asselmeyer, Sabine Bressler, Johannes Saukel, Gottfried Reznicek, and Christoph Dobeš. Quality control of valerianae radix by attenuated total reflection fourier transform infrared (atr-ftir) spectroscopy. *Planta Med*, 84(06/07):442–448, 26.04.2018. 442.

[74] Alexander Götz, Ramin Nikzad-Langerodi, Yannik Staedler, Anke Bellaire, and Johannes Saukel. Apparent penetration depth in attenuated total reflection fourier-transform infrared (atr-ftir) spectroscopy of allium cepa l. epidermis and cuticle. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 224:117460, 2020.

[75] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.