Proximal Methods for Calibration Transfer

Thomas Boucher*[1], M. Darby Dyar[2], and Sridhar Mahadevan[1]

[1]College of Information and Computer Sciences, University of Massachusetts,
Amherst, MA 01003 USA

[2]Department of Astronomy, Mount Holyoke College, South Hadley, MA 01075 USA

*boucher@cs.umass.edu

Summary

Abstract

Calibration transfer (CT) is the process of transferring a calibration curve from one instrument to another or from one set of conditions to another. Direct standardization (DS) of the spectra from a *source* to a *target* representation is a popular method of CT, but the multivariate objective function is often significantly underdetermined. Piecewise direct standardization regularizes DS by assuming only local differences between source and target spectra, but requires the same wavelength sampling between instruments. In this work a regularization framework from the field of convex optimization, proximal regularizers, is introduced to standardize instruments that sample at different wavelength ranges and where the differences may have global effects on the spectra. In this framework, penalty terms are appended to the DS objective function to enforce certain behaviors in the transfer matrix and the resulting transferred spectra, including sparsity and smoothness. This framework is shown to be effective at transferring spectra from a source NIR instrument with a narrow wavelength range to a target instrument with a much wider wavelength range. This is demonstrated using two publicly available NIR datasets.

KEYWORDS: calibration transfer, direct standardization, regularizers, proximal methods

Table of Contents Abstract

Proximal Methods for Calibration Transfer

Thomas Boucher, M. Darby Dyar, Sridhar Mahadevan

Abstract

A regularization framework for the calibration transfer method direct standardization (DS) is presented, *proximal DS*. In this framework, penalty terms are appended to the DS objective function to enforce certain behaviors in the transfer matrix and the resulting transferred spectra, including sparsity and smoothness. This framework is shown to be effective at transferring spectra from a source NIR instrument with a narrow wavelength range to a target instrument with a much wider wavelength range, where piecewise DS methods fail.

1 Introduction

In all spectroscopic applications, there is a need to ensure that possible differences in instrumentation, environment, or experimental conditions are mitigated. Calibration transfer (CT) is a technique for transferring a calibration curve from one instrument to another or from one set of environmental conditions to a differing set of conditions. CT can be performed by standardizing the model coefficients, the predicted values, or the spectral responses [1]. This work focuses on the latter, directly transferring spectra from one instrument's (or condition's) representation to another using a transfer function calculated from a small subset of standards recorded on both instruments. This method is known as direct standardization (DS) [2]. There are other CT methods that also operate directly with the spectral responses, but they function by transferring both instruments to a joint space [3] or to an instrument agnostic representation [4].

Perhaps the most popular alternative to DS is piecewise direct standardization (PDS) [2]. In this method, a series of piecewise functions are defined over windowed wavelength ranges to calculate a CT transfer map. When performing multivariate CT, the number of wavelength channels is often much larger than the number of standards. In this case, the DS problem benefits from regularization. PDS regularizes the DS problem by constraining the feature space to local wavelength neighborhoods. It performs well when transferring spectra with local differences and matching wavelength ranges [3].

In this work, we introduce a new regularization framework for DS based on recent advancements in the field of convex optimization, *proximal regularization*. In this approach, a single convex loss function is optimized that contains a series of penalty terms that encourage specific behaviors in the transfer function and in the resultant transferred spectra. By adding and removing penalty terms, a customized loss function is designed specifically for the data and task. Unlike PDS that optimizes many local loss functions, by optimizing a single loss function, *proximal DS* globally regularizes the transfer function and is able to correct for differences that span large wavelength regions, which can be caused in the spectra by matrix effects.

Some related work on regularization has been done in the field of calibration maintenance (CM). CM is a calibration transfer approach that modifies the calibration objective function instead of transferring the spectra. Regularized calibration maintenance has been discussed for l_1 and l_2 penalties [5, 6]. However, multiple algorithms were needed to optimize the different objective functions, and in this work, a single framework is presented to optimize many combinations of differentiable and non-differentiable penalties.

2 Background

Calibration transfer (CT) of the spectral responses is more formally defined as follows. Given a source set data $\mathcal{S} \in \mathbb{R}^{L\times p}$ and a target dataset $\mathcal{T} \in \mathbb{R}^{M\times q}$ of calibration spectra, where a small subset of N linking samples, $X \subset \mathcal{S}$ and $Y \subset \mathcal{T}$, have been recorded in both source and target formats, we seek to find a transfer function $f : \mathbb{R}^p \to \mathbb{R}^q$ such that,

$$f := \arg\min_f \frac{1}{2} \left\| f(X) - Y \right\|^2 . \tag{1}$$

The map f is referred to as the *transfer function*, and its purpose is to map samples from their source representation to their target representation. In practice, there are often too few calibration spectra X and Y to well-fit a non-linear transfer function.

In this work, we examine the case where f is a linear function, and so equation 1 can be rewritten using the linear transformation $T \in \mathbb{R}^{p\times q}$ as $f(X) = XT + B$, where B is a diagonal matrix bias term. Traditionally, this is solved using the pseudoinverse $T = U\Sigma^{-1}V^\top$, for the singular value decomposition, $\text{SVD}(T) = U\Sigma V^\top$, or more recent computationally efficient methods [7]. This method is called *direct standardization* (DS) in the chemometrics literature [2]. To get the most benefit from CT, it is desirable to have as few overlapping samples between source and target sets as possible. Unfortunately, this often results in a dramatically underdetermined system where $p \gg N$, which DS is has trouble solving. To help solve ill-posed problems like this, additional constraints can be added to the CT objective function (equation 1). These constraints, known as *regularizers*, often encourage traits, like smoothness and simplicity, or encode domain knowledge, like a structured feature space.

Perhaps the most popular regularized variant of direct standardization is the *piecewise direct standardization* (PDS) [2] method. PDS is a sliding window method that solves equation 1 by breaking the problem into q different least squares sub-problems, where each least squares problems uses a w-length window around the i^{th} channel of X to predict the i^{th} channel of Y. This results in $T^\top$ being a band matrix with w non-zero entries in each row. The piecewise nature of the solution has been shown to cause discontinuities in the resulting transferred spectra [8], so a regularized regression method must be used to solve the least squares problems to enforce a smooth result. In practice, partial least squares or principal component regression are used. PDS uses the structured (or ordered) nature of the feature representation to condition the CT problem by enforcing constraints on channel space. This method works well when the transfer instruments are similar, sharing the same wavelength range and sampling frequency; however, PDS cannot be used to transfer spectra of different wavelengths. For example, a spectrometer recording in the visual range cannot be transferred with PDS to a spectrometer in the near-infrared (NIR) range because the corresponding structured assumption of the spectra is no longer valid. Likewise, PDS cannot be used to transfer instruments with only overlapping wavelength regions. DS can be used in these more challenging transfer scenarios, but it must still be regularized.

3 Proximal Regularizers

PDS regularizes the problem by separating the CT problem into many sub-problems. However, equation 1 can also be regularized directly by appending a penalty to the transfer matrix T,

$$T := \arg\min_T \frac{1}{2} \left\| Y - XT - B \right\|^2 + \lambda g(T), \tag{2}$$

where g is a penalty function, often a norm $\|\cdot\|$ and $\lambda \geq 0$ is the penalty parameter. The function g is used to encourage certain properties in the solution T, like small-valued entries, that encourage different behaviors in the transferred spectra $XT + B$, like smoothness. For example, to encourage a sparse solution, where T has many zero entries, the l_1-norm can be used $g(T) = \|T\|_1 = \sum_{i=1}^{p} \sum_{j=1}^{q} |t_{i,j}|$, where $t_{i,j}$ is the i, j-element of the matrix T. This is an entry-wise matrix extension of the lasso model [9]. The l_1-norm is a case of the more general p-norm $\|T\|_p = \left(\sum_{i=1}^{p} \sum_{j=1}^{q} |t_{i,j}|^p \right)^{1/p}$. There are many benefits to a sparse solution, like decreased computing demands and increased interpretability. Many modern spectrometers record at hundreds or thousands of channels, which can result in a dense transfer matrix T with millions of entries. A sparse T greatly reduces the size on disk and in memory and leads to faster matrix computation. Also, by eliminating channels not used in the transfer function, T becomes human interpretable, allowing researchers to closely investigate the differences between source and target instruments or conditions.

Another entry-wise norm that can be used as a regularizing penalty is the *Frobenius norm*, also known as the Euclidean norm, $g(T) = \|T\|_F = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{q} |t_{i,j}|^2}$. Like the l_1 norm, the Euclidean norm is a p-norm where $p = 2$. Unlike the l_1 norm, the Euclidean norm does not induce sparsity, but rather shrinks the coefficients and smooths their values by drawing them into a similar range. In the residual, the penalty greatly decreases the variance, which improves the methods ability to transfer spectra outside of the original training set. Moreover, the smoothness in coefficients directly effects the smoothness of the resulting transferred spectra. As the penalty parameter λ is increased, the smoothness of the transferred spectra increases, within limits. This regularizer is commonly squared $\|\cdot\|_2^2$ to make it differentiable everywhere and is known as the ridge or Tikhonov regularizer.

To combine the sparsity of the lasso with the predictive power of the ridge, the elastic net $g(T) = \|T\|_1 + \frac{\gamma}{2} \|T\|_F^2$ was developed [10]. When the system is underdetermined $p < N$ and there are many collinear features, the typical scenario in spectroscopy, the performance of the lasso method is greatly improved with the addition of the ridge penalty. The presence of the l_1 norm will still encourage T to be sparse, but the l_2 term encourages smoothness in the transferred spectra. As γ is decreased, the elastic net solution will approach the lasso solution.

The regularizers discussed so far have been p-norms applied to matrices, but there is another family of matrix norms called Schatten norms that apply p-norms to the singular values of a matrix. For example, the Frobenius norm is also the $p = 2$ Schatten norm $\|T\|_F = \sqrt{\sum_{i=1}^{\min(p,q)} \sigma_i^2}$, where $\Sigma = \operatorname{diag}\left(\sigma_1, \ldots, \sigma_{\min(p,q)}, 0, \ldots, 0\right)$ are the singular values from the the singular value decomposition, $\mathrm{SVD}(T) = U\Sigma V^\top$. Another common regularizer is the Schatten norm $p = 1$, $\|T\|_* = \sum_{i=1}^{\min(p,q)} |\sigma_i|$, know as the *nuclear norm* (or trace norm). By applying the l_1 norm to the singular values of T, the nuclear norm forces the transfer matrix to have low rank. This low rank approximation of T is closely related to the principal components analysis (PCA) representation, as they both use the truncated SVD operation. In this way, the low rank regularizer encourages simplicity and reduces noise in the transfer function.

All of the regularizers g discussed have been norms, so the objective function in equation 2 is a convex function in all cases. Unfortunately, none of the regularizers are differentiable, so traditional gradient descent algorithms cannot be used. Instead, a class of algorithms known as *proximal methods* will be used. Proximal methods are general purpose convex optimization methods, but are especially well-suited to non-differentiable, penalized, large-scale problems [11]. Proximal algorithms use the *proximal operator* of a convex function g with domain $\mathcal{D}$,

$$\operatorname{prox}_{\lambda g}(x) = \operatorname*{argmin}_{u \in \mathcal{D}} \left(g(u) + \frac{1}{2\lambda} \|u - x\|_2^2 \right),$$

where $\lambda > 0$ is a mixing parameter controlling the difference penalty. The proximal operator (or mapping) can be interpreted as a generalization of the projection operator. If x is outside the domain $\mathcal{D}$ of g, then $\operatorname{prox}_{\lambda g}(x)$ will map x to a point in $\mathcal{D}$ that also minimizes g. Moreover, if g is the indicator function of a set $\mathcal{C}$, then $\operatorname{prox}_{\lambda g}(x)$ is the Euclidean projection onto $\mathcal{C}$.

The proximal operator can also be interpreted as a type of gradient method. The operator $\operatorname{prox}_{\lambda g}(x)$ yields the *proximal point* u that minimizes the function g while not straying too far from the point x. The parameter λ controls how far the proximal point can deviate from x, where higher values allow for greater deviation and stronger minimization of g. For a more complete listing of interpretations, reference chapter 3 of [11].

As a gradient method, the simplest proximal minimization algorithm is repeating

$$x^{k+1} = \operatorname{prox}_{\lambda g}\left(x^k\right), \tag{3}$$

where x^k indicates the point x during the k^{th} iteration of the algorithm, until $\left\|x^{k+1} - x^k\right\| < \epsilon$ for some very small $\epsilon > 0$. If g is a norm $\|\cdot\|$, then equation 3 will converge to the point x^* that minimizes g. [1] In the next section, the proximal operator for a few regularizing norms are discussed. Table I lists all of the proximal operators evaluated here.

The proximal operator of a vector norm $g = \|\cdot\|$ is $\operatorname{prox}_{\lambda g}(x) = x - \lambda \prod_{\mathcal{B}}(x/\lambda)$, where $\prod_{\mathcal{B}}$ is the projection onto the unit ball $\mathcal{B}$ of the norm. For a proof see section 6.5 of [11]. From this, the proximal operator for the l_1 norm, called *soft thresholding*, can be deduced. The soft thresholding operator is defined as

$$\operatorname{prox}_{\lambda\|\cdot\|_1}(T) = \begin{cases} T_{i,j} - \lambda, & T_{i,j} > \lambda \\ 0, & |T_{i,j}| \leq \lambda \\ T_{i,j} + \lambda, & T_{i,j} < -\lambda \end{cases} .$$

In some texts, this operator is also called the shrinkage operator, but we will follow the notation of [11] and reserve the latter term for the $\operatorname{prox}_{\|\cdot\|_F^2}$ operator.

Similarly, the proximal operator for the Frobenius norm, the entry-wise $\|\cdot\|_{l_2}$, is called the *block soft thresholding*,

$$\operatorname{prox}_{\lambda\|\cdot\|_F}(T) = \begin{cases} (1 - \lambda / \|T\|_F, & \|T\|_F \geq \lambda \\ 0, & \|T\|_F < \lambda \end{cases} .$$

The Frobenius norm is not differentiable everywhere, but the squared norm $\|\cdot\|_F^2$ is, so standard calculus can be used to derive its proximal operator, called the *shrinkage operator*, $\operatorname{prox}_{\lambda\|\cdot\|_F^2}(T) = \frac{1}{\lambda+1}T$. To calculate the proximal operator of the elastic net penalty, the soft

[1] This holds true more generally for all functions $g : \mathbb{R}^n \to \mathbb{R} \cup +\infty$ that are closed proper convex.

Regularizer	Objective	Proximal Operator	Algorithm
$\lambda \left\Vert T \right\Vert_1$	sparsity	soft thresholding	$\text{sign}(T) \cdot \max\left(0, \lvert T \rvert - \lambda\right)$
$\lambda \left\Vert T \right\Vert_2$	smoothness	block soft thresholding	$\max\left(0, \lvert T \rvert - \lambda\right) \frac{T}{\lVert T \rVert}$
$\lambda \left\Vert T \right\Vert_1 + \frac{\gamma}{2} \left\Vert T \right\Vert_2^2$	sparsity and smoothness	composition	$\frac{1}{1+\gamma}\text{sign}(T) \cdot \max\left(0, \lvert T \rvert - \lambda\right)$
$\lambda \left\Vert T \right\Vert_*$	low rank	singular value thresholding	$U \max\left(0, \Sigma - \lambda\right) V^\top$

Table I: Proximal regularizers used in experimentation, their primary objectives, and their one-line algorithms.

thresholding operator is composed with the shrinkage operator yielding,

$$\text{prox}_{\lambda \lVert \cdot \rVert_1 + \gamma \lVert \cdot \rVert_F^2}\left(T\right) = \frac{1}{\gamma + 1} \text{prox}_{\lambda \lVert \cdot \rVert_1}\left(T\right). \tag{4}$$

The nuclear norm $\left\Vert \cdot \right\Vert_*$ is a matrix norm, and so its proximal operator does not follow from the previous operators. However, it can be shown that the proximal operator of a Schatten p-norm is equal to the proximal operator of the vector p-norm applied to the singular values in the SVD decomposition. For $g = \left\Vert \cdot \right\Vert_*$ and $\text{SVD}(T) = U \Sigma V^\top$,

$$\text{prox}_{\lambda \lVert \cdot \rVert_*}(T) = U \left(\text{prox}_{\lambda \lVert \cdot \rVert_1}\left(\Sigma\right)\right) V^\top. \tag{5}$$

The same holds true for other Schatten norms $p \in [1, \infty]$, including the spectral norm $\left\Vert T \right\Vert_\infty = \max\left(\Sigma\right)$ whose proximal operator yields the best rank-1 approximation of T.

A complete listing of the proximal operator regularizers evaluated in section 4 is listed in Table I, along with simplified algorithms for their implementation. A large list of proximal operators can be found in Chapter 6 of [11].

To optimize equation 2 with a non-differentiable regularizer g, the proximal operator $\text{prox}_\lambda g$ can be used. To isolate the regularizer, the variable T can be split and the equation constrained,

$$\min_T \frac{1}{2} \left\Vert Y - XT \right\Vert^2 + \lambda g(Z) \text{ such that } T = Z. \tag{6}$$

The bias term B is omitted to decrease the notation because it is subtracted out before calculating T. While the splitting between T and Z may seem trivial, it allows the differentiable and non-differentiable terms of the objective function to be optimized separately. To impose the equality constraint $T = Z$ the *augmented Lagrangian* $\mathcal{L}_\rho$ is formed

$$\mathcal{L}_\rho(T, Z, Y) = \frac{1}{2} \|Y - XT\|^2 + g(Z) + \langle Y, T - Z\rangle + \frac{\rho}{2} \|T - Z\|^2, \tag{7}$$

where Y is the *Lagrange dual* variable to enforce the equality constraint and $\rho > 0$ is a penalty parameter controlling the rate of convergence by enforcing equality. The function $\mathcal{L}_\rho$ can be minimized using the constrained optimization algorithm *alternating direction method of multipliers* (ADMM) [12].

In a general form, the ADMM algorithm iterates over the three steps:

$$T^{k+1} = \arg\min_T \mathcal{L}_\rho(T, Z^k, Y^k) \tag{8}$$

$$Z^{k+1} = \arg\min_Z \mathcal{L}_\rho(T^{k+1}, Z, Y^k) \tag{9}$$

$$Y^{k+1} = Y^k + \rho(T^{k+1} - Z^{k+1}). \tag{10}$$

where T^k is the value for T at the k^{th} iteration, likewise for Z, Y. The first step (8) is a minimization of T. This can be solved in closed form using standard matrix calculus techniques, resulting in

$$T^{k+1} = \left(B^\top B + \rho I\right)^{-1} \left(B^\top A + \rho Z^k - Y^k\right).$$

The second step (9) is a minimization of Z. The regularizers studied in this work are non-differentiable, and so the proximal operator must be used in this step,

$$\begin{aligned} Z^{k+1} &= \arg\min_Z \left(g(Z) - trace\left(Y^{k\top} Z\right) + \frac{\rho}{2} \left\|T^{k+1} - Z\right\|_F^2\right) \\ &= \arg\min_Z \left(g(Z) + \frac{1}{2\lambda} \left\|T^{k+1} + \lambda Y^k - Z\right\|_F^2\right) \\ &= \operatorname{prox}_{\lambda g}\left(T^{k+1} + \lambda Y^k\right), \end{aligned}$$

setting $\lambda = 1/\rho$ and where $trace(\cdot)$ is the matrix trace. The last step (9) is an update of the Lagrange dual variable Y. These three steps are repeated until the variables Z and T converge. A simple test for convergence is if $\left\|Z^{k+1} - Z^k\right\| / \left\|Z^{k+1}\right\| < \epsilon_{tol}$ and $\left\|T^{k+1} - T^k\right\| / \left\|T^{k+1}\right\| < \epsilon_{tol}$ for some small tolerance like $\epsilon_{tol} = 10^{-4}$. Alternately, convergence of the primal residual $\left\|T^{k+1} - Z^{k+1}\right\|$ and the dual residual $\left\|-\rho(Z^{k+1} - Z^k)\right\|$ can be used for stopping criteria. See section 3.3.1 of [12] for greater detail.

In practice, ADMM typically converges quickly to a good solution. To decrease the computational burden of solving equation 8 in each iteration, the (symmetric, positive semi-definite) term $B^\top B + \rho I$ may be decomposed into triangular matrices using the Cholesky decomposition. Performing this operation once before gradient descent makes all subsequent calculations of T^{k+1} more efficient. For extremely large spectra with thousands of channels or more, it may be computationally advantageous to distribute the problem across channels. This is a detailed in section 8.3 of [12].

4 Experiments

Given a small training set of spectra recorded on both the source and target instruments, $\mathcal{S}_{train}, \mathcal{T}_{train}$, the task of the experiments was to learn a transfer function $T : \mathcal{S}_{train}T \approx \mathcal{T}_{train}$ to map a large testing set of spectra from their *source* representation $\mathcal{S}_{test}$ to their *target* representation $\mathcal{T}_{test}$. In these experiments, the source spectrometer had a narrow wavelength range, while the target spectrometer had a wavelength range many times larger than the source. In practice, using a technique like this a researcher could use a limited spectrometer with a very narrow wavelength range, like 200 nm, and transform it to a much wider spectrum, like 1400 nm, while sacrificing only a small amount of error. To evaluate the performance of the DS methods, the prediction error $\|\mathcal{S}_{test}T - \mathcal{T}_{test}\|_F$ and the relative prediction error $\|\mathcal{S}_{test}T - \mathcal{T}_{test}\|_F / \|\mathcal{T}_{test}\|_F$ were used to compare the predicted target test spectra with the actual target test spectra. All of the CT methods evaluated first mean centered the spectra, so a bias term has been omitted from the error description for simplicity. This error was used as a direct comparison of the transferred spectra. Instead of comparing the predictive accuracy of calibration models fit on the transferred spectra, which can confound the effectiveness of the DS methods with the choice of calibration model algorithm and parameter settings, this directly compared DS methods.

The evaluated CT methods were direct standardization (DS), elastic net DS, Euclidean DS, low rank DS, and sparse DS. A description of the proximal DS methods and their objective functions is listed in Table II. A Python implementation of all the methods evaluated is available through the authors website.[2] Two publicly available datasets of NIR spectra were used in the experimentation.[3] The performance of the CT methods were evaluated using varying training set sizes, where the training samples were subselected from a larger training set (from the source) using the Kennard-Stone selection method [13]. The sample closest (in Euclidean space) to the mean was first selected for the training set, then the set was constructed iteratively, where the next sample chosen was the one farthest from the closest current training set sample [14]. This method was used to select a representative source training subset.

4.1 NIR Corn Data

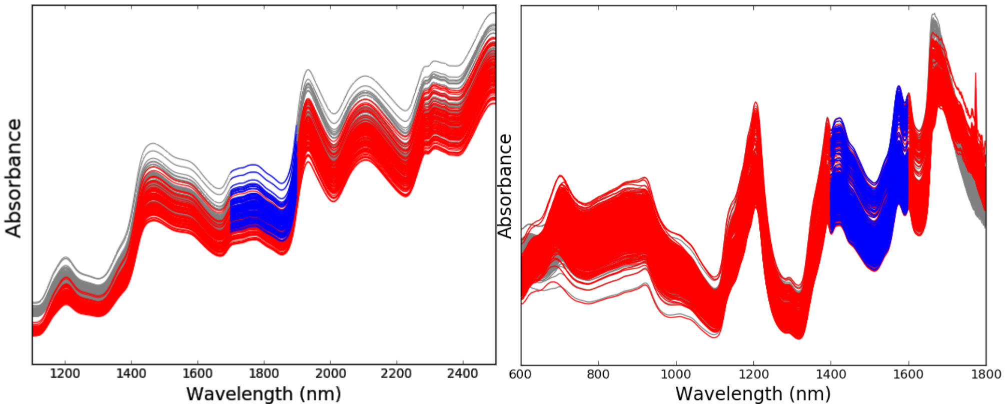
The dataset used in the first experiment was composed of 80 spectra of corn samples recorded using three different NIR spectrometers, labeled M5, MP5, and MP6, all from the same manufacturer and where MP5 and MP6 were located in the same facility. Each spectrum was recorded from 1100-2498 nm in 2 nm intervals resulting in 700 channels. This is a well studied dataset [15], but the experimental task conducted in this work was novel. Two sets of experiments were performed, one with M5 as the source and MP5 as the target and the other with M5 as the source and MP6 as the target. In both experiments, only wavelengths 1700-1898 nm (100 channels) of the source spectra were used. In figure 1 are example spectra from M5 and MP5 and the channels used during experimentation. To tune the CT methods, a 10 sample subset was selected from the source population using the Kennard-Stone method described above. For the proximal methods, leave-one-out cross validation was used to perform

[2] https://www.github.com/all-umass/

[3] http://www.eigenvector.com/data/

CT Method	Objective function	Comments
Sparse DS	$\min_T \frac{1}{2} \left\Vert Y - XT \right\Vert_F^2 + \alpha \left\Vert T \right\Vert_1$	Produces a sparse transfer function. More human interpretable. Especially advantageous for datasets with many channels. Resulting transferred spectra may be noisy.
Euclidean DS	$\min_T \frac{1}{2} \left\Vert Y - XT \right\Vert_F^2 + \alpha \left\Vert T \right\Vert_2$	Produces a smooth transfer function and transferred spectra. Like a non-differentiable ridge penalty.
ElasticNet DS	$\min_T \frac{1}{2} \left\Vert Y - XT \right\Vert_F^2 + \alpha \left\Vert T \right\Vert_1 + \frac{\gamma}{2} \left\Vert T \right\Vert_2^2$	A combination of the previous two regularizers, produces a sparse transfer matrix and smooth transferred spectra. Scales well for data with many channels.
Low Rank DS	$\min_T \frac{1}{2} \left\Vert Y - XT \right\Vert_F^2 + \alpha \left\Vert T \right\Vert_*$	Produces a low rank transfer function. Reduces noisy, unnecessary transforms in T using a PCA-like penalty.

Table II: The evaluated proximal calibration transfer (CT) methods, their loss functions, and some notes about each of the methods.



(a) Corn NIR spectra

(b) Tablet NIR spectra

Figure 1: The target spectra are red, the original full source spectra are in gray, and the source spectra used for experimentation are in blue.

a grid search over values for α and γ. In ADMM, for all proximal methods the convergence parameter $\lambda = 0.01$ was used. The other 70 samples were used to evaluated the CT methods using cross validation, where only a subset ($n = 10, 20, 30$) of the training samples were used to train the models during each fold. For each setting of n, cross validation was repeated 50 times with random shuffling between iterations.

	CT Method	$M5 \rightarrow MP5$	$M5 \rightarrow MP6$	Tablet
ElasticNet DS	Regularizer α	0.0001	0.001	0.0002
	Regularizer γ	0.003	0.003	0.0005
Euclidean DS	Regularizer α	0.5	0.4	0.01
Low Rank DS	Regularizer α	0.1	0.1	0.03
Sparse DS	Regularizer α	0.02	0.02	0.01

Table III: The parameter settings used by the calibration transfer techniques for the NIR experiments.

All proximal DS methods performed comparably well on the two tasks, $M5 \rightarrow MP5$ and $M5 \rightarrow MP6$ while standard DS performed significantly worse. The performance gap was most noticeable with fewer training examples. When $n = 10$, the error of standard DS was more than double any of the proximal DS methods. As the training set size n increased, the methods performance improved and became more equal. The prediction errors and the relative prediction errors and their standard errors for both tasks and all settings of n are listed in Table IV. Even though the small training set was selected to be representative of the larger set, the primary cause of failure for DS was still overfitting. In fact, since all of the proximal methods performed comparably, it suggests that many forms of regularization can remedy the problem.

	$M5 \rightarrow MP5$		
CT Method	$n = 10$	$n = 20$	$n = 30$
DS	1.91 ± 0.15 / 3.82% ± 0.30	1.52 ± 0.09 / 3.05% ± 0.18	1.39 ± 0.08 / 2.78% ± 0.16
ElasticNet DS	0.99 ± 0.03 / 1.99% ± 0.07	0.97 ± 0.03 / 1.94% ± 0.07	0.94 ± 0.03 / 1.88% ± 0.07
Euclidean DS	0.98 ± 0.04 / 1.97% ± 0.08	0.94 ± 0.03 / 1.88% ± 0.06	0.91 ± 0.04 / 1.82% ± 0.07
Low Rank DS	1.00 ± 0.04 / 2.01% ± 0.07	0.93 ± 0.03 / 1.86% ± 0.07	0.92 ± 0.03 / 1.85% ± 0.06
Sparse DS	1.00 ± 0.04 / 2.01% ± 0.08	0.97 ± 0.03 / 1.95% ± 0.06	0.97 ± 0.03 / 1.94% ± 0.07
	$M5 \rightarrow MP6$		
CT Method	$n = 10$	$n = 20$	$n = 30$
DS	1.93 ± 0.18 / 3.97% ± 0.36	1.56 ± 0.08 / 3.21% ± 0.17	1.47 ± 0.08 / 3.02% ± 0.17
ElasticNet DS	1.03 ± 0.05 / 2.13% ± 0.11	1.00 ± 0.05 / 2.06% ± 0.10	0.99 ± 0.05 / 2.04% ± 0.10
Euclidean DS	1.01 ± 0.06 / 2.09% ± 0.12	0.96 ± 0.05 / 1.98% ± 0.11	0.96 ± 0.05 / 1.98% ± 0.11
Low Rank DS	1.04 ± 0.05 / 2.14% ± 0.10	0.99 ± 0.05 / 2.03% ± 0.10	0.96 ± 0.05 / 1.98% ± 0.11
Sparse DS	1.03 ± 0.06 / 2.11% ± 0.12	1.01 ± 0.05 / 2.08% ± 0.12	1.00 ± 0.05 / 2.05% ± 0.10

Table IV: The prediction error / relative error ± the standard error of cross validation of NIR corn spectra setting M5 as the source and MP5 as the target on the top and setting M5 as the source and MP6 as the target on the bottom.

Although the prediction error of the proximal DS methods are similar, the transferred spectra they produced and the transfer functions were quite different. For example, the sparse DS model achieved error comparable to the other proximal methods, but the transferred spectra were extremely noisy. This was a direct result of the sparse transfer matrix. Shrinking entries of the transfer matrix to zero induced discontinuities in the resulting spectra. In the $M5 \rightarrow MP5$ experiment with $n = 10$, the sparse DS transfer matrix was $78.7\% \pm 0.01$ sparse, meaning that most of the entries in the matrix were zero. In figure 2 is a zoomed-in portion of a transferred spectrum showing the discontinuities imparted by sparse DS. In contrast, the Euclidean DS model produced smooth transferred spectra, but had a dense transfer matrix. To quantitatively measure the roughness of a set of transferred spectra $X^{N \times q}$, the sum of

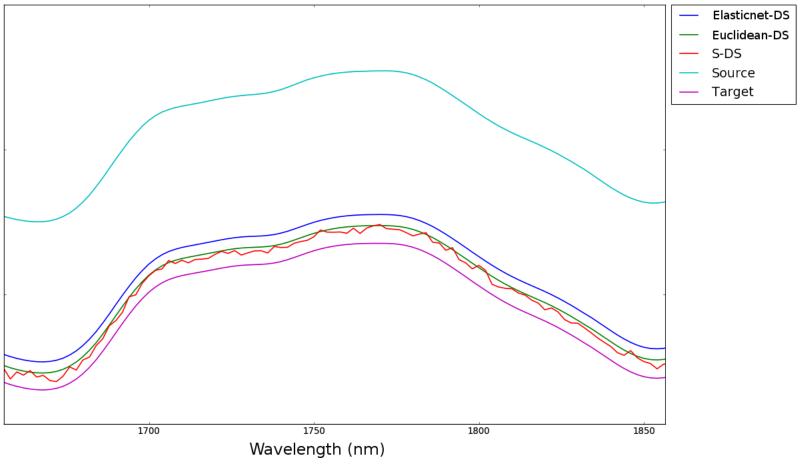


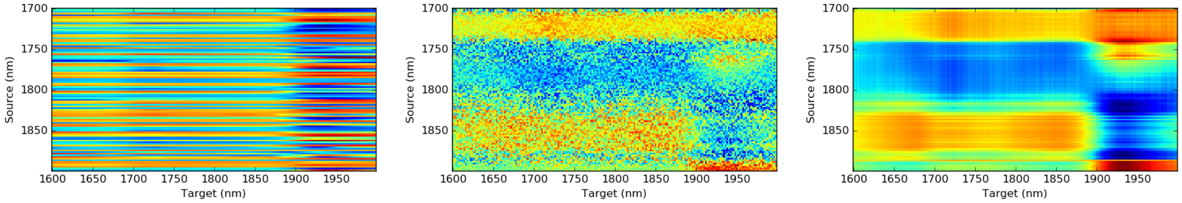
Figure 2: Applying Sparse DS, Euclidean DS, and ElasticNet DS to a test set spectrum from the NIR corn data. Sparse DS produces a sparse transfer matrix T, but T yields a rough spectrum. Euclidean DS yields a smooth spectrum, but has a dense T. ElasticNet DS is a combination of Sparse DS and squared Euclidean DS, and so yields a smooth spectrum with a sparse T.

squared second order differences was used,

$$\mathcal{R}(X) = \sum_{i=1}^{N} \sum_{j=2}^{q-1} \left((x_{i,j+1} - x_{i,j}) - (x_{i,j} - x_{i,j-1}) \right)^2 . \tag{11}$$

When X is perfectly smooth, $\mathcal{R}(X) = 0$, and as the roughness of X grows, $\mathcal{R}(x)$ increases. This is the same measure of smoothness used in Hodrick-Prescott filtering [16]. In the same experiment, the spectra from Euclidean DS had a roughness of $2.0 \times 10^{-3} \pm 2 \times 10^{-4}$, whereas sparse DS produced spectra with 3.75 times the roughness $7.5 \times 10^{-3} \pm 6 \times 10^{-4}$. In this case, the goal roughness of the target set was 1.5×10^{-3}. To smooth the spectra of sparse DS, an additional Euclidean penalty was added to form the elastic net DS model. The elastic net model was $64.53\% \pm 0.02$ sparse with a roughness of $2.9 \times 10^{-3} \pm 1 \times 10^{-4}$, a good compromise between sparsity and smoothness.

The low rank DS model produced transferred spectra similar to the Euclidean DS model. However, the transfer matrix of low rank DS was most similar to standard DS, because instead of directly shrinking the transfer matrix entries it operated on the components of the matrix. In figure 3 are portions of the transfer matrices for standard DS, Euclidean DS, and low rank



(a) DS

(b) Euclidean DS

(c) Low Rank DS

Figure 3: A portion of the transfer matrix for three of the direct standardization (DS) methods, (a) DS, (b) Euclidean DS, (c) Low Rank DS. All 100 source channels are displayed, but for clarity only 300 target channels are displayed. Observe that (a) overfits the data, having large variance between rows, while (b) and (c) have been regularized and are more generalizable. (b) appears more pixelated than (c) because its regularization is local to the matrix entries, whereas (c) regularizes the matrix globally by shrinking its singular values.

DS. The pixelated appearance of the Euclidean DS matrix was caused by the local entry-wise regularization, versus the smooth gradation of the low rank DS matrix that globally regularized by shrinking the singular values.

In these experiments, only a small portion (14%) of the available channels from the source spectra were used to predict the entire target spectra. Even with this handicap, and using just ten training samples, the transferred target spectra from the proximal methods had on average only a 1.99% error (for $M5 \rightarrow MP5$). For a comparison with ideal conditions, the experiment was repeated using all of the source channels. In this case, piecewise models like PDS and block-style proximal methods could be used and were included in the experiment. Even with this expanded list of competing DS methods and access to the full wavelength source spectra, the top model still had an average error of 1.83%. This small difference in accuracy of just 0.16% indicates the feasibility of this method of using proximal DS to extrapolate large wavelength regions from much smaller regions.

4.2 NIR Tablet Data

In the second experiment, the same task was repeated using a different NIR dataset to ensure the results were reproducible across datasets, predicting the full target spectra from a partial source spectra. The experimental dataset was composed of two subsets, a 154 sample calibration set and a 459 sample held-out test set. Each spectrum was recorded from 600-1898 nm in 2 nm intervals resulting in 650 channels. The last 50 channels of both spectrometers were nearly entirely noise, so these channels were omitted from the experiment. Only 100 channels were used in the source spectra, from 1400-1598 nm. Figure 1 contains examples of the source and target spectra and the utilized wavelength ranges. To tune the proximal method parameters, cross validation grid search over the calibration set was used. For each validation fold, the Kennard-Stone algorithm was used to select 15 training samples. The final tuned parameter settings are listed in table III. In ADMM, for all proximal methods the convergence parameter $\lambda = 0.9$ was used. There was little difference in performance as λ varied, so the setting was selected to keep the number of ADMM iterations low. After the parameters were tuned for each method, they were fitted using the calibration set and evaluated on the held-out test set.

The prediction error and relative prediction error over the test set are listed in table V. The size of the training set was varied $n = 10, 15, 20, 25$. When the training set size was small $n = 10$, all of the proximal DS methods greatly outperformed standard DS, which had twice the error of any proximal method. As the training set size increased, the performance gap between the proximal methods and standard DS decreased.

Similar to corn experiments, there was little difference in the predictive error of the proximal methods. Overall, Euclidean DS performed slightly better than the competing proximal methods. However, its performance worsened for the largest setting of $n = 25$. This could be due to statically setting $n = 15$ during the parameter tuning, and refitting the parameters for each training set size would likely eliminate this spike in prediction error.

In ideal unhindered conditions, using all 650 source channels, the global proximal methods outperformed the local piecewise DS methods. With $n = 15$, Euclidean DS achieved a relative error of 0.82% whereas the best piecewise method had 1.01% error. The difference between using the full source versus the partial source was 0.86% relative error. While this was still small compared to the fraction of channels used in the hindered experiment, this error was larger than the 0.16% difference reported in the first experiment experiment. This is likely due to how well the source channel subsets represent the instrument differences. Moreover, global differences in the tablet set may reach beyond the small 100 channel window.

CT Method	$n = 10$	$n = 15$	$n = 20$	$n = 25$
DS	79.40 / 4.02%	51.40 / 2.60%	37.97 / 1.92%	33.75 / 1.71%
ElasticNet	36.86 / 1.87%	34.05 / 1.72%	33.04 / 1.67%	32.35 / 1.64%
Euclidean DS	36.35 / 1.84%	33.26 / 1.68%	32.27 / 1.63%	32.85 / 1.66%
Low Rank DS	36.54 / 1.85%	33.60 / 1.70%	32.55 / 1.65%	32.74 / 1.66%
Sparse DS	36.50 / 1.85%	33.51 / 1.70%	32.48 / 1.64%	32.66 / 1.65%

Table V: The prediction error / relative prediction error over the held-out NIR tablet test set.

5 Conclusion

In this work, a new method for regularizing direct standardization (DS) is presented, proximal DS. Using proximal methods from the field of optimization, non-differentiable convex penalty terms are added to DS to enforce certain characteristics, like sparsity or smoothness, and to prevent overfitting. Proximal DS is especially useful for solving standardization tasks between instruments of varying wavelength regions, like inferring a complete NIR spectra from only a portion. This is shown experimentally using two well studied NIR spectra datasets. With a NIR corn dataset, all proximal DS methods were shown to transfer spectra with around 2% relative error using only 10 training standards, which is only .2% more relative error than the best performing method using the entire source wavelength. Similar results were shown for the NIR tablet dataset.

Only four proximal DS methods were evaluated during experimentation, but many more are easily derived. Furthermore, many of the penalties can be combined to form new methods, For example, for a robust low rank DS method, the error can be directly modeled using a sparse term $\|T\|_* + \|E\|_1$. The conditions of the CT task and the type of spectroscopy may require different penalty schemes. In future work, we plan to specifically investigate proximal DS methods for laser-induced breakdown spectroscopy (LIBS), where matrix effects in the plasma can cause global discrepancies between the source and target instruments that locally regularized methods cannot correct.

6 Acknowledgments

We are grateful for support from NSF grants CHE-1306133 and CHE-1307179.

References

(1) Feudale, R. N.; Woody, N. A.; Tan, H.; Myles, A. J.; Brown, S. D.; Ferré, J. *Chemometr. Intell. Lab.* **2002**, *64*, 181–192.

(2) Wang, Y.; Veltkamp, D. J.; Kowalski, B. R. *Anal. Chem.* **1991**, *63*, 2750–2756.

(3) Walczak, B; Bouveresse, E; Massart, D. *Chemometr. Intell. Lab.* **1997**, *36*, 41–51.

(4) Geladi, P; MacDougall, D; Martens, H *Appl. Spectrosc.* **1985**, *39*, 491–500.

(5) Kalivas, J. *J. Chemometrics* **2012**, 218–230.

(6) Kalivas, J. H.; Siano, G. G.; Andries, E.; Goicoechea, H. C. *Appl. Spectrosc.* **2009**, *63*, 800–809.

(7) Drineas, P.; Mahoney, M. W.; Muthukrishnan, S; Sarlós, T. *Numerische Mathematik* **2011**, *117*, 219–249.

(8) Gemperline, P. J.; Cho, J.; Aldridge, P. K.; Sekulic, S. S. *Anal. Chem.* **1996**, *68*, 2913–2915.

(9) Tibshirani, R. *J. Royal Stat. Soc., Series B* **1994**, *58*, 267–288.

(10) Zou, H.; Hastie, T. *J. Royal Stat. Soc., Series B* **2005**, *67*, 301–320.

(11) Parikh, N.; Boyd, S. *Found. Trends Optim.* **2014**, *1*, 127–239.

(12) Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. *Found. Trends Mach. Learn.* **Jan. 2011**, *3*, 1–122.

(13) Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137–148.

(14) Daszykowski, M; Walczak, B; Massart, D. *Anal. Chim. Acta* **2002**, *468*, 91–103.

(15) Wang, Y.; Kowalski, B. R. *Appl. Spectrosc.* **1992**, *46*, 764–771.

(16) Hodrick, R.; Prescott, E. *J. Money, Credit and Banking* **1997**, *29*, 1–16.