Preprints of the 10th IFAC International Symposium on Dynamics and Control of Process Systems
The International Federation of Automatic Control
December 18-20, 2013. Mumbai, India

Spectroscopic monitoring of diesel fuels using Supervised Distance Preserving Projections

Francesco Corona [*] **Zhanxing Zhu** [**] **Amauri H. Souza Júnior** [***]
Michela Mulas [****] **Roberto Baratti** [†]

[*] *Department of Information and Computer Science, Aalto University, School of Science, Espoo, Finland (e-mail: francesco.corona@aalto.fi)*
[**] *Institute for Neural and Adaptive Computation, School of Informatics, University of Edinburgh, Edinburgh, UK (e-mail: zhanxing.zhu@ed.ac.uk)*
[***] *Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza, Brazil (e-mail: amauriholanda@ifce.edu.br)*
[****] *Department of Civil and Environmental Engineering, Aalto University, School of Engineering, Espoo, Finland (e-mail: michela.mulas@aalto.fi)*
[†] *Department of Mechanical, Chemical and Material Engineering, University of Cagliari, Cagliari, Italy (e-mail: baratti@dicm.unica.it)*

Abstract: In this work, we discuss a recently proposed approach for supervised dimensionality reduction, the Supervised Distance Preserving Projection and, we investigate its applicability to monitoring material's properties from spectroscopic observations. Motivated by continuity preservation, the SDPP is a linear projection method where the local geometry of the points in the low-dimensional subspace mimics the geometry of the points in the response space. Such a mapping facilitates an efficient regressor design and it may also uncover useful information for visualisation. An experimental evaluation is conducted to show the performance of the SDPP and compare it with a number of state-of-the-art approaches for unsupervised and supervised dimensionality reduction. For the task, the results obtained on a benchmark problem consisting of a set of NIR spectra of diesel fuels and six different chemico-physical properties of those fuels is discussed. Based on the experimental results, the SDPP leads to accurate and parsimonious projections that can be used in the design of efficient regression models.

Keywords: Supervised Distance Preserving Projection, Dimensionality reduction, Principal Component Analysis, Partial Least Squares, Multivariate quality control, Soft-sensor design

1. INTRODUCTION

Spectrophotograms are recognised sources of information in a variety of fields ranging from analytical chemistry to process industry. Many applications reported in the research and industrial literature regard the estimation of important quality indexes (typically, chemical and physical properties) in a material starting from a collection of light absorbance spectra (Workman, 1999). The information encoded in the spectra results from the interaction between light and matter and it is displayed as complex curves conditioned by the composition of the analysed samples. The composition, in turn, determines the properties of interest. Without specific methods of analysis, such information is not easily accessible and, cannot be directly extracted and used for estimation purposes. In fact, one intrinsic characteristic of the measurements acquired by a spectrophotometer is that the absorbance spectrum can be regarded as a regular function observed at discretised arguments in the instrument's operating range of wavelengths. Because of such a distinctive feature, the calibration problem of estimating the response output (the property of interest) is defined from very high-dimensional and collinear input covariates (the spectra).

To address the calibration problem, one common regression approach is used in practice. The standard solution is to rely on full-spectrum methods for linear dimensionality reduction coupled with linear regression. Reference models and *de facto* standard in multivariate calibration are the well-known Principal Component Regression, which performs Principal Component Analysis (PCA, Jolliffe (2002)) followed by Multiple Linear Regression (MLR), and Partial Least-Squares Regression, which combines Projection to Latent Structures (PLS, Wold et al. (2001)) and MLR. PCA is an unsupervised dimensionality reduction method that learns a low-dimensional input subspace by maximising the variance of the covariates and PLS is a supervised method that constructs a low-dimensional input subspace by maximising the covariance between the projected covariates and the output. Following the advances in dimensionality reduction, the *kernelised* extensions Kernel-PCA (KPCA, Schölkopf et al. (1998)) and Kernel-PLS (KPLS, Rosipal and Trejo (2002)) have been developed and used to perform nonlinear projections of the spectral data (Rosipal et al., 2003).

In this work, we discuss a recently proposed approach for supervised dimensionality reduction, the Supervised Distance Preserving Projection (SDPP, Zhu et al. (2013)) and we investigate its applicability to the calibration problem from spectroscopic observations. Motivated by continuity preservation, the SDPP minimises the difference between distances among projected covariates and distances among responses, locally. The minimisation of distance differences leads to the effect that the geometry of the input points in the low-dimensional subspace mimics the geometry of the corresponding points in the response space.

Being the projection map linear and parametric, the SDPP can easily handle the out-of-sample data. Such a simple map facilitates the design of efficient regressors that estimate product's properties from very high-dimensional spectral observations.

The remainder of this paper is organised as follows. Section 2 overviews the Supervised Distance Preserving Projection and the two optimisation schemes to solve it. In Section 3, an experimental evaluation is conducted to show the performance of the SDPP and compare it with four state-of-the-art approaches: PCA, PLS, KPCA and KPLS. For the task, a benchmark problem from the Southwest Research Institute (SwRI) consisting of a set of Near Infrared (NIR) spectra of diesel fuels and six different chemico-physical properties of those fuels is discussed.

2. THE SDPP

The Supervised Distance Preserving Projection (SDPP) is dimensionality reduction method based on simple geometric intuitions on the assumed continuity of the mapping from the covariates to the response space. The Weierstrass definition of continuity of a function states that if two points are close in the covariates space, then they are also close in the response space; The SDPP is designed to find a low-dimensional subspace where such a continuity is preserved. In the following, the basic formulation of the SDPP is overviewed, along with the two principal optimisation schemes designed to solve it.

2.1 Formulation of the SDPP

Formally, we are given n data points $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\} \in \mathbb{R}^d$ and their corresponding responses $\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n\} \in \mathbb{R}^m$, and we assume the existence of a continuous mapping $f : \mathscr{X} \mapsto \mathscr{Y}$. Provided that the input space $\mathscr{X}$ is well-sampled, we expect that for each point $\mathbf{x} \in \mathscr{X}$ and for every $\varepsilon_\mathbf{y} > 0$ there exists an $\varepsilon_\mathbf{x} > 0$ such that $d(\mathbf{x},\mathbf{x}') < \varepsilon_\mathbf{x} \Rightarrow \delta(f(\mathbf{x}),f(\mathbf{x}')) < \varepsilon_\mathbf{y}$, where $d(\cdot,\cdot)$ and $\delta(\cdot,\cdot)$ are distance functions in $\mathscr{X}$ and $\mathscr{Y}$, respectively. Under this condition, the Supervised Distance Preserving Projection is designed to compute a low-dimensional subspace $\mathscr{Z}$ of dimensionality r with $r \ll d$, where such a continuity is preserved. The SDPP achieves this by *matching* the local geometry of the data points in the $\mathscr{Z}$ and $\mathscr{Y}$ spaces. The geometrical structure is expressed by pairwise distances over neighbourhoods of the input covariates. Inside the neighbourhoods, the SDPP minimises the difference between distances among projected covariates and distances among responses.

The Supervised Distance Preserving Projection assumes that the subspace $\mathscr{Z}$ can be obtained by a linear transformation of $\mathscr{X}$; that is, for an input point $\mathbf{x}$, the new representation in the subspace is $\mathbf{z} = \mathbf{W}^T\mathbf{x}$, where the projection matrix $\mathbf{W} \in \mathbb{R}^{d\times r}$. Concretely, the SDPP seeks for a linear transformation $\mathbf{W}$ that parameterises the input distances by minimising the criterion

$$J(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathbf{x}_j\in\mathscr{N}(\mathbf{x}_i)}(d_{ij}^2(\mathbf{W}) - \delta_{ij}^2)^2, \qquad (1)$$

where $\mathscr{N}(\mathbf{x}_i)$ is a neighbourhood of $\mathbf{x}_i$. To characterise pairwise distances, the conventional Euclidean metric is commonly used; that is, $d_{ij}^2(\mathbf{W}) = \|\mathbf{z}_i - \mathbf{z}_j\|^2$ and $\delta_{ij}^2 = \|\mathbf{y}_i - \mathbf{y}_j\|^2$.

Figure 1 pictorially depicts the functioning of the SDPP, where, for an input point $\mathbf{x}$, three nearest neighbours $\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}$ are considered and a transformation $\mathbf{W}$ that leads to a similar geometry between the $\mathscr{Z}$-space and the $\mathscr{Y}$-space is found. To match the local geometry of the $\mathscr{Y}$-space, one of the three nearest neighbours, $\mathbf{x}_2$, is *moved*, after projection, outside the neighbourhood in the $\mathscr{Z}$-space while another point is *moved* inside. This match is beneficial to the regression from the subspace $\mathscr{Z}$ to the response space $\mathscr{Y}$ and to the visualisation of the relationship existing between inputs and responses.

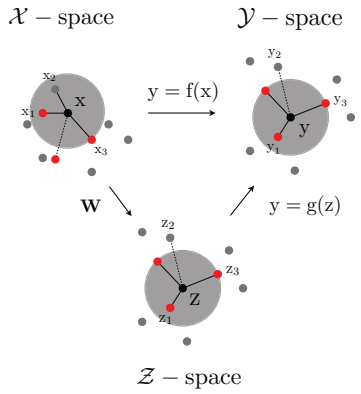

Fig. 1. A schematic illustration of the SDPP. Solid lines indicate connections between nearest neighbours.

The criterion of the SDPP, Equation 1, is similar to the S-Stress (Takane et al., 1977), the objective used in one of the variants of another dimensionality reduction approach, the Multidimensional Scaling (MDS, Cox and Cox (2000)). Both the SDPP and the S-Stress are, in fact, formulated from *squared* pairwise distances, whereas the Kruskal's Stress used in other MDS variants is based on plain pairwise distances. The S-Stress is, however, defined for an unsupervised method and, thus no response information is used. Moreover, the S-Stress pursues global distance preservation, whereas the SDPP captures the local geometry by incorporating a neighbourhood graph into the cost. In the formulation of the Supervised Distance Preserving Projection, locality for each data point $\mathbf{x}_i$ is explicitly controlled by considering its k nearest neighbours $\mathscr{N}(\mathbf{x}_i)$. The number of neighbours k is thus a hyper-parameter of the SDPP and it has to be provided by the user beforehand or to be tuned from data.

2.2 Optimisation of the SDPP

To optimise the objective function of the Supervised Distance Preserving Projection, two different strategies have been designed: i) a Semidefinite Quadratic Linear Programming (SQLP) problem and ii) a Conjugate-Gradient (CG) optimisation. The two formulations are overviewed in the following.

SQLP Starting from the square of the pairwise distances

$$\begin{aligned} d_{ij}^2(\mathbf{W}) &= (\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{W}\mathbf{W}^T(\mathbf{x}_i-\mathbf{x}_j) \\ &= (\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{P}(\mathbf{x}_i-\mathbf{x}_j), \end{aligned}$$

with $\mathbf{P} = \mathbf{W}\mathbf{W}^T$ a positive semidefinite (PSD) matrix denoted as $\mathbf{P} \succeq 0$, the optimisation of SDPP can be formulated as an instance of *convex quadratic semidefinite programming* (QSDP).

After defining $\boldsymbol{\tau}_{ij} = \mathbf{x}_i - \mathbf{x}_j$, the squared pairwise distances (parameterised by the projection model $\mathbf{W}$) can be written as

$$d_{ij}^2(\mathbf{W}) = \boldsymbol{\tau}_{ij}^T\mathbf{P}\boldsymbol{\tau}_{ij} = \mathtt{vec}(\boldsymbol{\tau}_{ij}\boldsymbol{\tau}_{ij}^T)^T\mathtt{vec}(\mathbf{P}) = \mathbf{l}_{ij}^T\mathbf{p},$$

where vector $\mathbf{l}_{ij} = \texttt{vec}(\tau_{ij}\tau_{ij}^T)$ and vector $\mathbf{p} = \texttt{vec}(\mathbf{P})$. $\texttt{vec}(\cdot)$ is an operator that concatenates all the columns of a matrix into a new vector. The objective can be re-written as a function of $\mathbf{p}$,

$$\begin{aligned} J(\mathbf{p}) &= \mathbf{p}^T \underbrace{\left(\frac{1}{n}\sum_{ij}\mathbf{G}_{ij}\mathbf{l}_{ij}\mathbf{l}_{ij}^T\right)}_{\mathbf{A}}\mathbf{p} + \underbrace{\left(-\frac{2}{n}\sum_{ij}\mathbf{G}_{ij}\delta_{ij}^2\mathbf{l}_{ij}\right)^T}_{\mathbf{b}}\mathbf{p} \\ &\quad + \underbrace{\frac{1}{n}\sum_{ij}\mathbf{G}_{ij}\delta_{ij}^4}_{c} \\ &= \mathbf{p}^T\mathbf{A}\mathbf{p} + \mathbf{b}^T\mathbf{p} + c, \end{aligned} \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{d^2\times d^2}, \mathbf{b} \in \mathbb{R}^{d^2\times 1}$, and c is a constant that can be ignored later in the optimisation. In Equation 2, $\mathbf{G}_{ij}$ is used to denote the neighbourhood graph of $\mathbf{x}_i$. $\mathbf{G}_{ij}$ is defined as

$$\mathbf{G}_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_j \text{ is a neighbor of } \mathbf{x}_i, \\ 0 & \text{otherwise.} \end{cases}$$

The SDPP is then optimised from the equivalent QSDP problem

$$\begin{aligned} &\min_{\mathbf{p}} \ \mathbf{p}^T\mathbf{A}\mathbf{p} + \mathbf{b}^T\mathbf{p} \\ &\text{s.t. } \mathbf{P} \succeq 0 \end{aligned} \tag{3}$$

It is important to notice that the QSDP formulation does not optimise the projection matrix $\mathbf{W}$ directly, instead it optimises the PSD matrix $\mathbf{P} = \mathbf{W}\mathbf{W}^T$. The projection matrix $\mathbf{W}$ can be computed either as the square root of $\mathbf{P}$ or, alternatively, from a Singular Value Decomposition of $\mathbf{P}$ to obtain an orthogonal projection matrix $\mathbf{W}$. In the latter case, the i-th column of $\mathbf{W}$ is calculated as $\sqrt{\lambda_i}\mathbf{v}_i$, being λ_i and $\mathbf{v}_i$ the i-th eigenvalue and eigenvector of $\mathbf{P}$, respectively. The dimensionality of the projection subspace is determined by analyzing the eigenvalues.

Equation 3 can be also written as a *semidefinite programming* (SDP) problem. Because of the low-rank structure of $\mathbf{A}$, the QSDP problem can be in fact reformulated into the equivalent *semidefinite quadratic linear programming* (SQLP) problem

$$\begin{aligned} &\min_{\mathbf{p},\mathbf{u}} \ (\mathbf{e}_1 - \mathbf{e}_2)^T\mathbf{u} + \mathbf{b}^T\mathbf{p} \\ &\text{s.t. } (\mathbf{e}_1 + \mathbf{e}_2)^T\mathbf{u} = 1, \\ &\quad \mathbf{B}\mathbf{p} - \mathbf{C}\mathbf{u} = \mathbf{0}, \\ &\quad \mathbf{u} \in \mathbb{K}_{q+2}, \\ &\quad \mathbf{P} \succeq 0, \end{aligned} \tag{4}$$

where q is the rank of $\mathbf{A}$, by Cholesky factorisation $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ with $\mathbf{B} \in \mathbb{R}^{q\times d^2}$, $\mathbf{C} = [0_{q\times 2}, \mathbf{I}_{q\times q}]$ and $\mathbf{e}_i$ is the i-th basis vector with $i = 1, 2, \ldots, q+2$. $\mathbb{K}_m$ is used to denote the second-order cone of dimension m (i.e., $\mathbb{K}_m = \{(x_0; \mathbf{x}) \in \mathbb{R}^m | x_0 \geq \|\mathbf{x}\|\}$). The SQLP formulation sets the problem into a standard framework of optimisation with semidefinite constraints, which is supported by many efficient optimisation libraries.

The solution of the SQLP problem requires $O(d^{6.5})$ arithmetic operations and it is, in that sense, convenient when compared to the SDP solution that requires $O(d^9)$ operations.

Conjugate-Gradient optimisation When the dimensionality of the original dataset is very high, the size of $\mathbf{A}$ in the SQLP formulation becomes extremely large. Due to the large size of the matrix $\mathbf{A}$, the SQLP solution is therefore feasible only for not very high-dimensional problems (e.g. when $d < 100$). This aspect brings practical limitations related to storing capacity and further optimisation. To overcome these shortcomings, an alternative optimisation approach based on the conventional *conjugate-gradient* (CG) search has been formulated.

Algorithm 1 Conjugate-Gradient optimisation of SDPP

Input: Training data matrices $\mathbf{X}$ and $\mathbf{Y}$, neighbourhood graph $\mathbf{G}$, initialised projection matrix $\mathbf{W}_0$
Output: Optimised projection matrix $\mathbf{W}$.

1. Compute gradient $\nabla_{\mathbf{W}}J$;
2. Vectorize the projection matrix, $\mathbf{w}_0 = \texttt{vec}(\mathbf{W}_0)$;
3. Vectorize the gradient, $\mathbf{g}_0 = \texttt{vec}(\nabla_{\mathbf{W}}J)$;
4. Initialize the conjugate direction as $\nu_0 = -\mathbf{g}_0$;
for $t = 1 \rightarrow T$ **do**
  5. Calculate β_t by Polak-Ribiére's rule, $\beta_t = \frac{\mathbf{g}_t^T(\mathbf{g}_t - \mathbf{g}_{t-1})}{\mathbf{g}_{t-1}^T\mathbf{g}_{t-1}}$;
  6. Update the conjugate direction, $\nu_t = -\mathbf{g}_t + \beta_t\nu_{t-1}$;
  7. Perform line search, $\eta_t = \arg\min_\eta \ J(\mathbf{w} + \eta\nu_t)$;
  8. Update $\mathbf{w}$, $\mathbf{w}_{t+1} = \mathbf{w}_t + \eta_t\nu_t$
end for
9. Reshape the vector $\mathbf{w}_{T+1}$ into the matrix $\mathbf{W}$.

After denoting the (squared) pairwise distances as $\mathbf{D}_{ij} = d_{ij}^2(\mathbf{W})$ and $\Delta_{ij} = \delta_{ij}^2$, the objective function in Equation 1 is written as

$$J(\mathbf{W}) = \frac{1}{n}\sum_{ij}\mathbf{G}_{ij}(\mathbf{D}_{ij} - \Delta_{ij})^2 \tag{5}$$

The gradient with respect to $\mathbf{W}$ is then equal to $\nabla_{\mathbf{W}}J = 4/n\sum_{ij}\mathbf{G}_{ij}(\mathbf{D}_{ij} - \Delta_{ij})\tau_{ij}\tau_{ij}^T\mathbf{W}$. A more compact form of the gradient can be obtained after denoting $\mathbf{Q} = \mathbf{G} \odot (\mathbf{D} - \Delta)$ with $\odot$ representing the element-wise product of two matrices, the symmetric matrix $\mathbf{R} = \mathbf{Q} + \mathbf{Q}^T$ and $\mathbf{S}$ a diagonal matrix with $\mathbf{S}_{ii} = \sum_j \mathbf{R}_{ij}$. Straightforward algebraic manipulations lead to

$$\nabla_{\mathbf{W}}J = \frac{4}{n}\mathbf{X}^T(\mathbf{S} - \mathbf{R})\mathbf{X}\mathbf{W}, \tag{6}$$

where each row of the data matrix $\mathbf{X}$ is a data point $\mathbf{x}_i$ and $\mathbf{L} = \mathbf{S} - \mathbf{R}$ is the Laplacian matrix. The *conjugate-gradient* optimisation of the objective of the SDPP is given in Algorithm1.

It is worth noticing that the CG approach allows for a direct optimisation of the projection matrix $\mathbf{W}$. In comparison to the SQLP approach where the dimensionality of the projection subspace is selected *a posteriori*, here it is defined beforehand.

3. MONITORING DIESEL FUELS

In this section, we illustrate the effectiveness of the Supervised Distance Preserving Projection and we compare its performance with four state-of-the-art methods for unsupervised and supervised dimensionality reduction. For comparison, we consider Principal Component Analysis (PCA), Partial Least Squares (PLS), Kernel Principal Component Analysis (KPCA) and Kernel Partial Least Squares (KPLS). When kernel methods are used, standard Gaussian kernels are employed, with the optimal kernel width estimated by cross-validation. As for the neighbourhood size in SDPP, the heuristic to define locality to be equal to 10% of the available data points is used ($k = 0.1n$).

The application consists of analysing six different properties in summer diesel fuels starting from a set of spectral observations. The data are provided by the Southwest Research Institute (`http://www.swri.org`) and publicly available for benchmarking purposes from the Eigenvector Research Incorporated (`http://www.eigenvector.com`). The absorbance spectra are acquired by means of a spectrophotometer operating in the $900 - 1700$nm range. The absorbance is measured on

the basis of the NIR principle with a 2nm resolution, Figure 2. Each observation consists of the 401-channel spectrum of absorbances ($\mathbf{x}_i \in \mathbb{R}^d$, with $d = 401$) and the corresponding values of six different chemico-physical properties ($\mathbf{y}_i \in \mathbb{R}^m$, with $m = 6$): i) Boiling point at 50% recovery; ii) Cetane number; iii) Density at 15°C; iv) Freezing temperature; v) Total Aromatics; and, vi) Viscosity at 40°C. The measurements of the product's properties are obtained in laboratory by reference methods. The dataset consists of $n = 135$ observations for learning the projection models and 125 observations for testing. The six outputs and corresponding spectra are analysed independently.

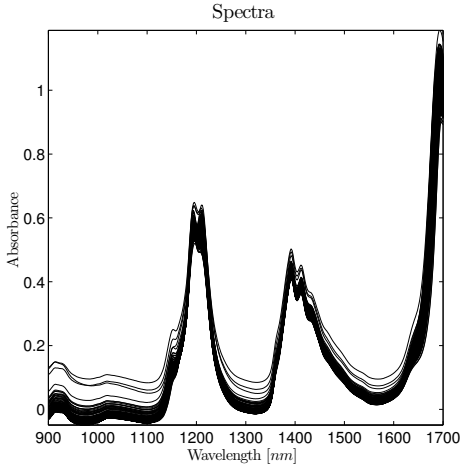

Fig. 2. Visualisation of the spectra in the wavelength domain.

The spectra show the typical overlapped absorbance bands arising from different hydrocarbon functional groups and reflect the samples composition. The major absorbance features in the experimental region are assigned to the second overtone ($\approx 1100-1300$nm), the combination bands ($\approx 1300-1550$nm) and the first overtone ($\approx 1600-1800$nm) of the Carbon-Hydrogen vibrations (Wheeler, 1959; Weyer, 1985). In details, in the second overtone region, we can observe the aromatic bonds at $\approx$ 1150nm, the methylene bonds at $\approx$ 1220nm indicating the presence of linear hydrocarbons, whereas the methyl bonds at $\approx$ 1200nm indicate the presence of branched hydrocarbon although the absorbance is also influenced by the amount of linear paraffin. By the same token, the combination bands for methylene ($\approx$ 1400nm), and methyl ($\approx$ 1350nm) mimic what observed in the short-wavelength range. The vibrations of the C-H bond on different functional groups lead to distinct absorption peaks, therefore, the six chemico-physical properties of the fuels can be explored from spectra since phenomenological relationships between chemical structure and properties exist.

3.1 Visualisation of the projected spectra

In order to get an insight on the spectra and their low-dimensional arrangement with respect to the six properties, we projected the spectra ($d = 401$) onto a bi-dimensional subspace ($r = 2$). For the task only the learning points were used. The testing points were projected afterwards with the out-of-sample formulations of the methods. The 2D subspace was selected to support the presentation on easily intelligible visual displays.

Figure 3 shows the bi-dimensional projections of the input spectra using a colouring scheme that dyes the points according to the corresponding values of the response, for each property and each for each method. From the figure, it is possible to notice how the projections obtained with supervised methods (PLS and KPLS) appear visually superior when compared to what obtained with the unsupervised methods (PCA and KPCA); an expected result. On the other hand, the bi-dimensional subspaces learned by the SDPP are based on two highly informative features that allow for an ordered arrangement of the projected input points with respect to the responses. This is particularly true for the density, the total aromatics and the viscosity of the fuel samples. For such properties, the input spectra are arranged almost linearly, indicating that a mono-dimensional projection would be sufficient for reconstructing the outputs. For the cetane number and the freezing point, also for the SDPP projections onto a higher a number of features seem to be needed for reconstructing the responses.

The qualitative assessment of the projections and corresponding visualisations can be quantified after recalling that when the data dimensionality is reduced it is not necessarily possible to preserve all the similarities. The reduction causes two main kind of errors: i) Data point that are not neighbours in the original space can be mapped close by in the projection space, causing data points to be falsely identified as similar; and, ii) Data points that are neighbours in the original space can be mapped far away in the projection space, causing similarity relations not to be correctly reconstructed. Based on Venna and Kaski (2001), such errors can be used to measure the *trustworthiness* and the *continuity* of the $\mathscr{X} \to \mathscr{Z}$ mapping.

- The *trustworthiness* of a projection is defined by denoting with $U_{k_r}(i)$ the set of points that are in k_r-neighbourhood of $\mathbf{z}_i$ in the projection space but not in the original one and, with $\tilde{r}(i,j)$ the rank of $\mathbf{x}_j$ in the ordering based on its distance from $\mathbf{x}_i$. Trustworthiness of $\mathscr{X} \to \mathscr{Z}$ is then
$$M_{\text{trust}}^{\mathscr{X}\to\mathscr{Z}}(k_r) = 1 - C(k_r)\sum_{i=1}^{n}\sum_{j\in U_{k_r}(i)}(\tilde{r}(i,j)-k_r)$$
- The *continuity* of a projection is defined by letting $V_{k_r}(i)$ be the set of points that are in the k_r-neighbourhood of $\mathbf{x}_i$ in the $\mathscr{X}$-space but not in the $\mathscr{Z}$-space and, by letting $\hat{r}(i,j)$ be the rank of $\mathbf{z}_j$ in the ordering based on its distance from $\mathbf{z}_i$. Continuity of $\mathscr{X} \to \mathscr{Z}$ is then
$$M_{\text{cont}}^{\mathscr{X}\to\mathscr{Z}}(k_r) = 1 - C(k_r)\sum_{i=1}^{n}\sum_{j\in V_{k_r}(i)}(\hat{r}(i,j)-k_r),$$

The neighbourhood size k_r is to be understood as the amplitude of the region of interest over which the figures of merit are evaluated. The term $C(k_r)$ simply scales the measures into $[0,1]$:
$$C(k_r) = \begin{cases} \dfrac{2}{nk_r(2n-3k_r-1)} & \text{if } k_r < \dfrac{n}{2}, \\ \dfrac{2}{n(n-k_r)(n-k_r-1)} & \text{if } k_r \geq \dfrac{n}{2}. \end{cases}$$

Figure 4 shows the measures of trustworthiness and continuity of the bi-dimensional projections achieved by PCA, PLS, KPCA, KPLS and SDPP for a region of interest k_r ranging in $[2,64]$. The diagrams highlight how PCA and PLS are the best performers, with both trustworthiness and continuity monotonically increasing with the amplitude of the region of interest. This is not surprising considering that PCA can be understood as a method for globally preserving pairwise distances and PLS is known to find features that are often similar to the principal components. On the other hand, the corresponding kernel extensions returned projections that are only moderately faithful. Similar results are also obtained by the SDPP, indicating that

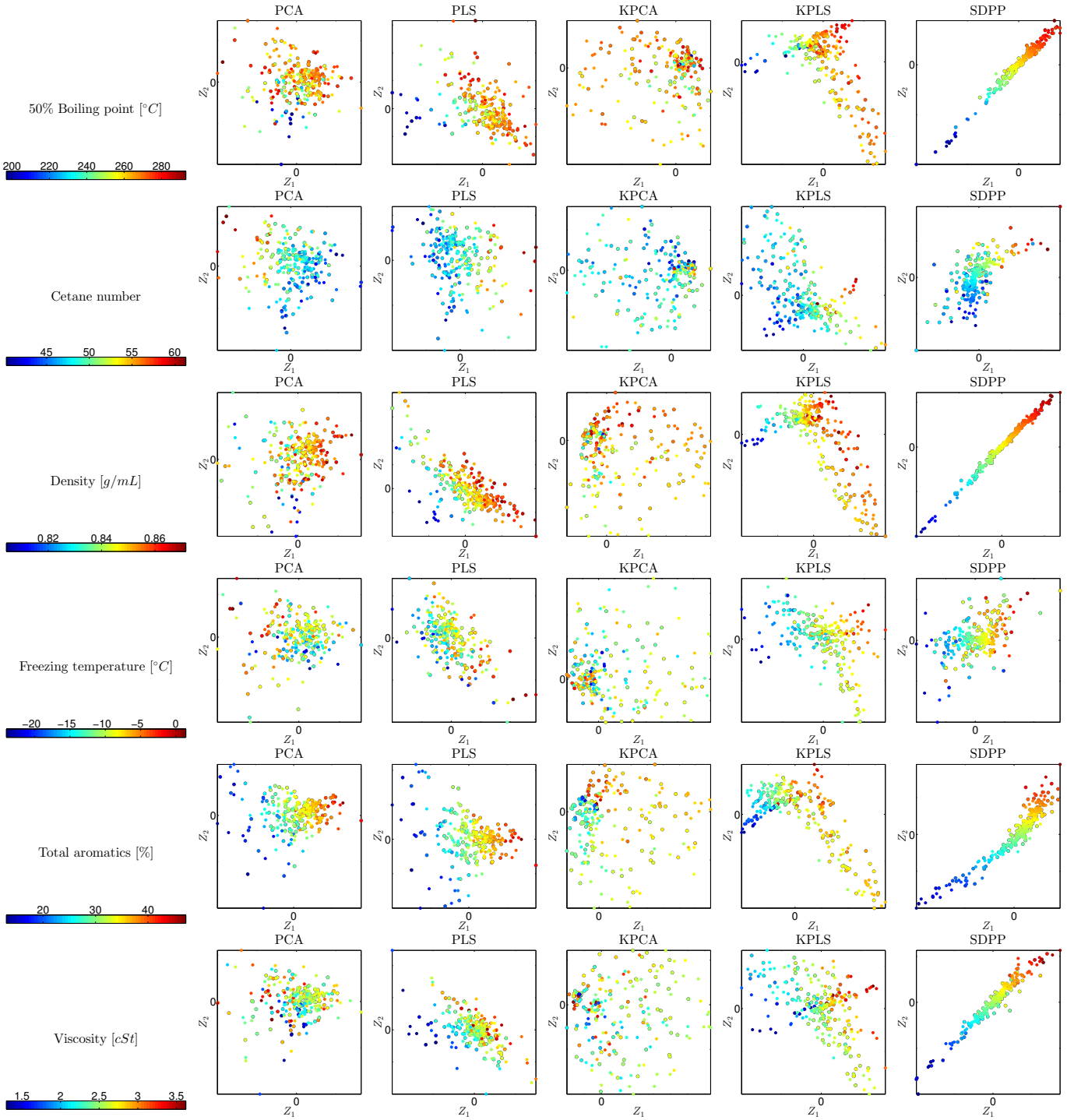

Fig. 3. Bi-dimensional projection and visualisation of the input spectra. Colouring based on output values is used to dye the inputs.

the apparent quality of its bi-dimensional displays does not necessarily correspond to an accurate preservation of the similarities existing between spectra, before and after projection.

This result is not surprising because such criterion is not embedded in the SDPP's cost function. In that respect, the SDPP aims at mapping inputs characterised by similar outputs close by in the projection space. To design accurate regressors it is, in fact, more desirable and expected that the continuity of the mapping from $\mathscr{X}$ to $\mathscr{Y}$ is as high as possible. In the spirit of Venna and Kaski (2001), a measure of the continuity of the $\mathscr{X} \to \mathscr{Y}$ map can be defined by letting $V_{k_r}(i)$ be now the set of points that are in the neighbourhood of size k_r in the $\mathscr{X}$-space but not in the $\mathscr{Y}$-space and, by letting $r(i,j)$ be the rank of y_j in the ordering based on its distance from y_i. The continuity of $\mathscr{X} \to \mathscr{Y}$ is

$$M_{\text{cont}}^{\mathscr{X}\to\mathscr{Y}}(k_r) = 1 - C(k_r) \sum_{i=1}^{n} \sum_{j \in V_{k_r}(i)} (r(i,j) - k),$$

Note that here k_r and $C(k_r)$ bear the same meaning as before, whereas k is the locality parameter of the SDPP (i.e., $k = 0.1n$).

Figure 5 shows the measure of continuity for regression after the bi-dimensional projections achieved by PCA, PLS, KPCA, KPLS and the SDPP. Again, a region of interest k_r ranging in

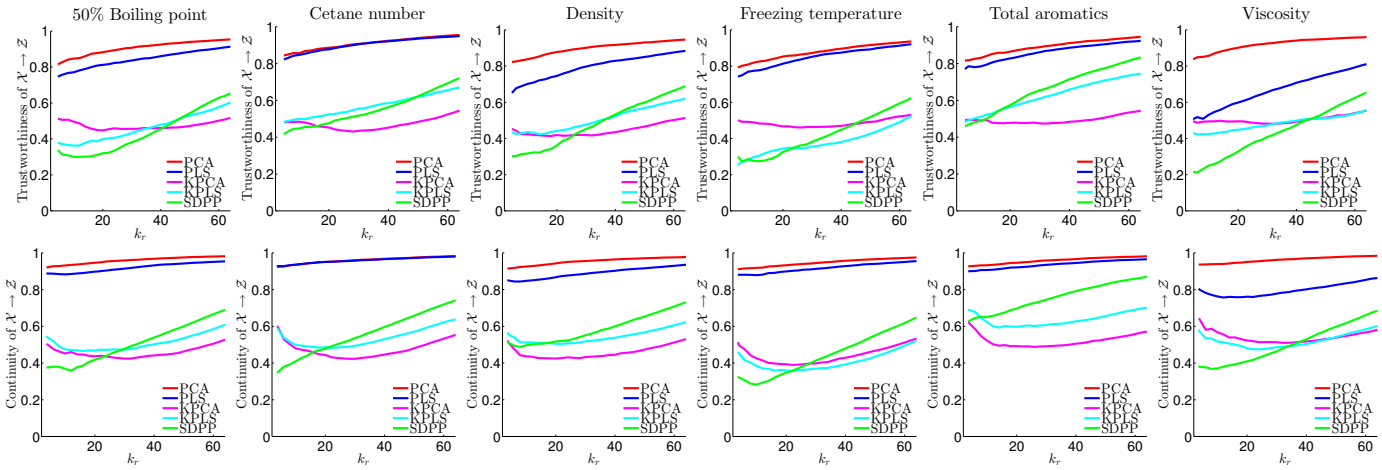

Fig. 4. Trustworthiness and continuity of the $\mathscr{X} \to \mathscr{Z}$ projection and visualisation, for a region of interest $k_r \in [2,64]$.

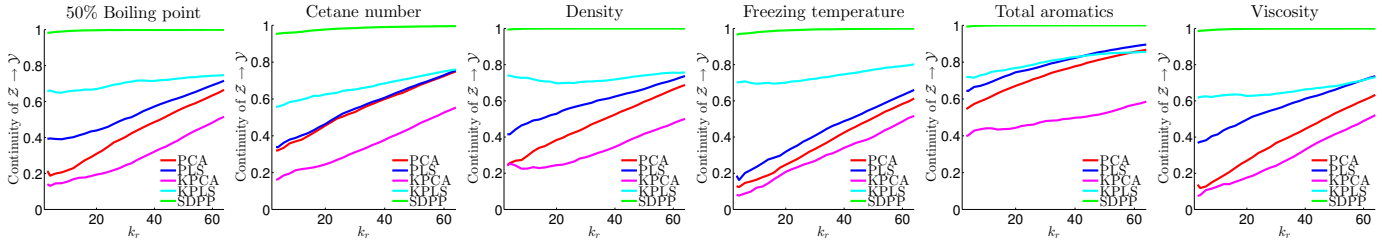

Fig. 5. The measure of continuity for the $\mathscr{Z} \to \mathscr{Y}$ regression, for a region of interest $k_r \in [2,64]$.

$[2,64]$ is used. In this case, the diagrams highlight how SDPP is consistently the best performer in accurately representing the continuity between the projected spectra and their corresponding responses, for a wide amplitude of the region of interest. This is also true for those outputs that appeared to require a higher number of features to faithfully capture the input-output relationships. The result suggests that, for all the responses, simple linear regression models calibrated globally over all the learning samples projected by the SDPP would be sufficient to estimate the properties of the fuels from the projected spectra.

4. CONCLUSIONS

The Supervised Distance Preserving Projection is a supervised dimensionality reduction method designed to project high-dimensional covariates onto a low-dimensional subspace where the geometry of the input points mimics the geometry of the corresponding output points in the response space. Such type of projection is desirable for designing accurate and yet parsimonious regression models from very high dimensional and possibly correlated input spaces. This type of regression problems is typically encountered in chemometrics, where the calibration of material's properties from very high-dimensional spectral observations remains a major application area. In this work, the applicability of the SDPP under these ill-posed regression conditions is investigated on a set of NIR spectra of diesel fuel samples and six corresponding chemico-physical properties. Based on the experimental results, we found that the SDPP can be used to generate informative and yet parsimonious projections finalised to the design of efficient calibration models.

REFERENCES

Cox, T.F. and Cox, M.A.A. (2000). *Multidimensional Scaling, Second Edition*. Chapman and Hall/CRC.

Jolliffe, I.T. (2002). *Principal Component Analysis, Second Edition*. Springer.

Rosipal, R. and Trejo, L.J. (2002). Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Mach. Learn. Res.*, 2, 97–123.

Rosipal, R., Trejo, L.J., Matthews, B., and Wheeler, K. (2003). Nonlinear kernel-based chemometric tools: A machine learning approach. In *International Symposium on PLS*, 249–260.

Schölkopf, B., Smola, A., and Müller, K.R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5), 1299–1319.

Takane, Y., Young, F.W., and Leeuw, J.D. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42(1), 7–67.

Venna, J. and Kaski, S. (2001). Neighborhood preservation in nonlinear projection methods: An experimental study. In *International Conference on Artificial Neural Networks*, 485–491.

Weyer, L.G. (1985). Near-infrared spectroscopy of organic substances. *Appl. Spectrosc. Reviews*, 21(1-2), 1–43.

Wheeler, O.H. (1959). Near infrared spectra of organic compounds. *Chem. Rev.*, 59(4), 629–666.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab.*, 58(2), 109–130.

Workman, J.J.J. (1999). Review of process and non-invasive near-infrared and infrared spectroscopy: 1993-1999. *Appl. Spectrosc. Reviews*, 34, 1–89.

Zhu, Z., Similä, T., and Corona, F. (2013). Supervised distance preserving projections. *Neural Process. Lett.*, In Press.