Monitoring diesel fuels with Supervised Distance Preserving Projections and Local Linear Regression

Francesco Corona *, Zhanxing Zhu †, Amauri H. Souza Júnior ‡,
Michela Mulas §, Guilherme A. Barreto ‡ and Roberto Baratti ¶
* Department of Information and Computer Science, Aalto University, Espoo, Finland.
E-mail: francesco.corona@aalto.fi
† Institute for Neural and Adaptive Computation, 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, guilherme@deti.ufc.br
§ Department of Civil and Environmental Engineering, Aalto University, 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 (SDPP) and, we investigate its applicability to monitoring material's properties from spectroscopic observations using Local Linear Regression (LLR). An experimental evaluation is conducted to show the performance of the SDPP and LLR 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 are discussed. Based on the experimental results, the SDPP leads to accurate and parsimonious projections that can be effectively used in the design of estimation models based on local linear regression.**

I. 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 from a collection of light absorbance spectra [1].

The information encoded in the spectra results from the interaction between light and matter and it is observed as complex curves conditioned by the composition of the analysed samples. The composition, in turn, determines the property of interest. Without specific methods of analysis, such information is not easily accessible and, cannot be directly extracted and used in estimation tasks. 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). Furthermore, it is not unusual to analyse datasets with a number of observations that is radically smaller than the number of input covariates.

To address this ill-conditioned 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 (PCR), which performs Principal Component Analysis (PCA, [2]) followed by Multiple Linear Regression (MLR), and Partial Least-Squares Regression (PLSR), which combines Projection to Latent Structures (PLS, [3]) 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, kernel extensions like Kernel-PCA (KPCA, [4]) and Kernel-PLS (KPLS, [5]) have been developed and used to firstly perform a nonlinear projection of the spectral data and then regress the output.

In this work, we discuss a recently proposed approach for supervised dimensionality reduction, the Supervised Distance Preserving Projection (SDPP, [6]). Specifically, we investigate the applicability of the SDPP to the calibration problem from spectroscopic observations when it is coupled with Local Linear Regression (LLR, [7]). 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. LLR are ensembles of regression models calibrated only on small subsets of input observations in a neighbourhood which is similar to the new inputs. Usually, the local regressors are simple linear models like the aforementioned MLR. Similarity in LLR is conventionally based on distance and, either on a fixed number of *nearest neighbours* or on a varying number of neighbours that adapt to the local topology of the data, as with convex neighbourhoods like *natural neighbours*, *natural neighbours inclusive* and *enclosing k-nearest neighbours* ([8], [9]). Local

978-1-4799-3194-1/13 $31.00 © 2013 IEEE
DOI 10.1109/BRICS-CCI-CBIC.2013.76

linear regressors are globally nonlinear, they can achieve high accuracies and be updated to automatically include new points.

The remainder of this paper is organised as follows. Section II overviews the SDPP and Section III presents LLR and techniques for neighbourhood definition. In Section IV, an experimental evaluation is conducted to show the performance of the SDPP coupled with LLR and compare it with four state-of-the-art approaches (PCA, PLS, KPCA and KPLS). A benchmark problem from the Southwest Research Institute consisting of a set of Near Infrared (NIR) spectra of diesel fuels and six different properties of those fuels is discussed.

II. Supervised Distance Preserving Projections

The Supervised Distance Preserving Projection (SDPP) is a 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 formulation of the SDPP and its optimisation is overviewed.

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 computes 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

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

where $\mathscr{J}_{\mathbf{x}_i}$ 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 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

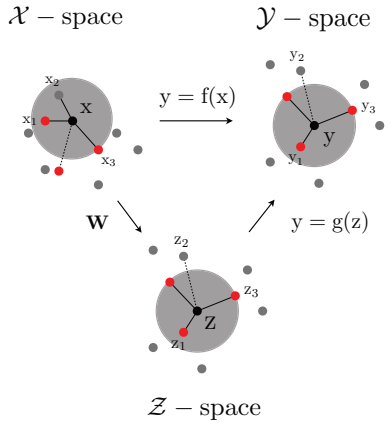


Fig. 1. The SDPP: Solid lines denote connections between nearest neighbours.

subspace $\mathscr{Z}$ to the response space $\mathscr{Y}$ and to the visualisation of the relationship existing between inputs and responses.

A. 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 $d_{ij}^2(\mathbf{W})=(\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{W}\mathbf{W}^T(\mathbf{x}_i-\mathbf{x}_j)=(\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{P}(\mathbf{x}_i-\mathbf{x}_j)$, with $\mathbf{P}=\mathbf{W}\mathbf{W}^T$ a positive semidefinite matrix $\mathbf{P}\succeq 0$, 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 $\mathbf{W}$) are 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 $\mathbf{l}_{ij}=\mathtt{vec}(\boldsymbol{\tau}_{ij}\boldsymbol{\tau}_{ij}^T)$ and $\mathbf{p}=\mathtt{vec}(\mathbf{P})$. The $\mathtt{vec}(\cdot)$ operator concatenates the columns of a matrix into a vector. Then, 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. $\mathbf{G}_{ij}$ denotes the neighbourhood graph of $\mathbf{x}_i$ and it is defined in such a way that $\mathbf{G}_{ij}=1$, if $\mathbf{x}_j$ is a neighbour of $\mathbf{x}_i$, and $\mathbf{G}_{ij}=0$ otherwise.

The SDPP is optimised from the equivalent QSDP problem

$$\min_{\mathbf{p}}\ \mathbf{p}^T\mathbf{A}\mathbf{p}+\mathbf{b}^T\mathbf{p},\quad\text{s.t.}\quad\mathbf{P}\succeq 0. \tag{3}$$

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 (Cholesky Decomposition) of $\mathbf{P}$ or, alternatively, from a Singular Value Decomposition of $\mathbf{P}$ to obtain an orthogonal matrix $\mathbf{W}$.

Equation 3 can be written also as a *semidefinite programming* (SDP) problem. In this case, the QSDP problem is reformulated into a *semidefinite quadratic linear programming* (SQLP) problem that conveniently requires $O(d^{6.5})$ arithmetic operations, whereas the SDP solution needs $O(d^9)$ operations.

Conjugate-Gradient optimisation: When the dimensionality d is very high, the size of $\mathbf{A}$ in the SQLP formulation becomes extremely large. 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 *conjugate-gradient* (CG) search has been formulated.

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

$$J(\mathbf{W})=\frac{1}{n}\sum_{ij}\mathbf{G}_{ij}(\mathbf{D}_{ij}-\Delta_{ij})^2 \qquad (4)$$

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 $\nabla_{\mathbf{W}}J=\frac{4}{n}\mathbf{X}^T(\mathbf{S}-\mathbf{R})\mathbf{X}\mathbf{W}$. Each row of $\mathbf{X}$ is a data point $\mathbf{x}_i$ and $\mathbf{L}=\mathbf{S}-\mathbf{R}$ is the Laplacian matrix.

Note 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.

III. Local Linear Regression

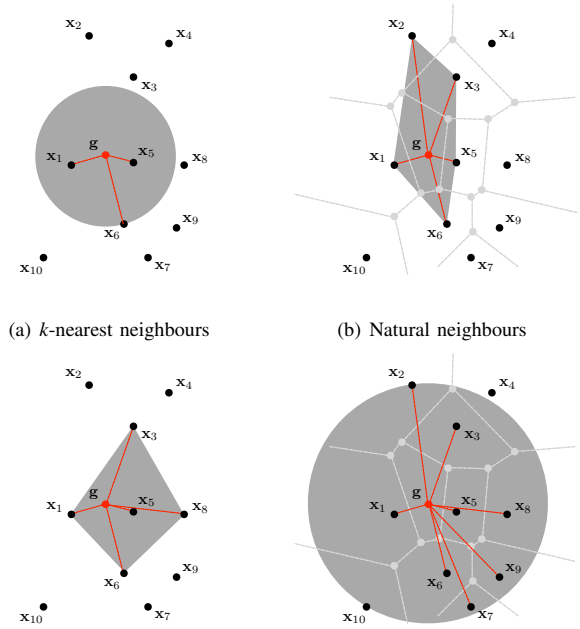
Local Linear Regression (LLR) is a nonlinear estimation approach. The spirit of LLR is that, over a small subset of the input domain, a simple MLR model can approximate sufficiently well the true mapping to the output. LLR retains the simplicity of MLR and it can overcome its low accuracy.

We are given n training points $\mathscr{X}\rightarrow\mathscr{Y}=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$, where $\mathbf{x}_i\in\mathbb{R}^d$ and $y_i\in\mathbb{R}$. For an arbitrary input test point $\mathbf{g}\in\mathbb{R}^d$, LLR estimates its output as $\hat{y}=\hat{\beta}^T\mathbf{g}+\hat{\beta}_0$, the least-squares hyperplane over the neighbourhood $\mathscr{J}_\mathbf{g}$ of $\mathbf{g}$:

$$(\hat{\beta},\hat{\beta}_0)=\underset{\beta,\beta_0}{\operatorname{argmin}}\sum_{\mathbf{x}_j\in\mathscr{J}_\mathbf{g}}(y_i-\beta^T\mathbf{g}-\beta_0)^2.$$

The definition of the neighbourhood and the number of neighbours are crucial for local linear regression. In this section, we briefly define and illustrate four major neighbourhood definition strategies for LLR, from a geometrical point of view.

Classic *k-nearest neighbours* (kNN) define a neighbourhood $\mathscr{J}_\mathbf{g}^{kNN}$ of $\mathbf{g}$ using k of its neighbours, according to a specified distance metric. Usually, the Euclidean metric is used and the number of neighbours k is fixed or cross-validated.



(a) k-nearest neighbours (b) Natural neighbours

(c) Enclosing k-nearest neighbours (d) Natural neighbours inclusive

Fig. 2. Neighbourhoods: a) $\mathscr{J}_\mathbf{g}^{kNN}=\{\mathbf{x}_1,\mathbf{x}_5,\mathbf{x}_6\}$, b) $\mathscr{J}_\mathbf{g}^{NN}=\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_5,\mathbf{x}_6\}$, c) $\mathscr{J}_\mathbf{g}^{ekNN}=\{\mathbf{x}_1,\mathbf{x}_3,\mathbf{x}_5,\mathbf{x}_6,\mathbf{x}_8\}$ and d) $\mathscr{J}_\mathbf{g}^{NNi}=\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_5,\mathbf{x}_6,\mathbf{x}_7,\mathbf{x}_8,\mathbf{x}_9\}$.

Figure 2(a) shows a kNN neighbourhood of size $k=3$ for a test point $\mathbf{g}$. Despite its simplicity, one major problem in kNN is the selection of the neighbourhood size: i) too few neighbours may lead to a neighbourhood that does not enclose the test point which might give a large estimation variance and, ii) too many neighbours to impose enclosure may cause the model to over-smooth. How to select adaptively k is an open issue.

A. Enclosing neighbourhoods

If $\mathscr{J}_\mathbf{g}$ encloses $\mathbf{g}$, we call it an enclosing neighbourhood; i.e., $\mathbf{g}\in\texttt{conv}(\mathscr{J}_\mathbf{g})$, where the convex hull of a point set $S=\{\mathbf{s}_1,\ldots,\mathbf{s}_n\}$ is defined as $\texttt{conv}(S)=\{\sum_{i=1}^n\omega_i\mathbf{s}_i|\sum_{i=1}^n\omega_i=1,\omega_i\geq 0\}$. Recently, [8] proved that if a test point is in the convex hull enclosing its neighbourhood, then the variance of the local linear regression estimate is bounded by the variance of the measurement noise. Such a property is fundamental to limit erratic results. In the following, three enclosing neighbourhood definition strategies are briefly overviewed.

Enclosing k-nearest neighbours (ekNN): It is based on the kNNs of $\mathbf{g}$ and extends them to define a neighbourhood that encloses it, Figure 2(c). ekNN is the neighbourhood of the kNNs with the smallest k such that $\mathbf{g}\in\texttt{conv}(\mathscr{J}_\mathbf{g}(k))$, where $\mathscr{J}_\mathbf{g}(k)$ is the set of kNNs of $\mathbf{g}$ [8]. If $\mathbf{g}$ is outside of convex hull of the set $\mathscr{X}$, no such k exists. Define *distance to enclosure* as $D(\mathbf{g},\mathscr{J}_\mathbf{g})=\min_{\mathbf{z}\in\texttt{conv}(\mathscr{J}_\mathbf{g})}\|\mathbf{g}-\mathbf{z}\|_2$, where $\mathbf{z}$ is any point in the convex hull around the neighbourhood of $\mathbf{g}$. Note that $D(\mathbf{g},\mathscr{J}_\mathbf{g})=0$ only if $\mathbf{g}\in\texttt{conv}(\mathscr{J}_\mathbf{g})$. Then, the e$k$NN neighbourhood is $\mathscr{J}_\mathbf{g}(k^*)$ with $k^*=\min_k\{k|D(\mathbf{g},\mathscr{J}_\mathbf{g}(k))=D(\mathbf{g},\mathscr{X})\}$. The complexity for building a convex hull using k neighbours is $O(k^{\lfloor d/2\rfloor})$, where $\lfloor\cdot\rfloor$ is the floor function.

Natural neighbours (NN): Natural neighbours are based on the Voronoi tessellation of the training samples and the test point. The natural neighbours of $\mathbf{g}$ are defined as those points whose Voronoi cells are adjacent to the cell including $\mathbf{g}$. Natural neighbours have the so-called *local coordinates property*, which is used to prove that the natural neighbours form an enclosing neighbourhood if $\mathbf{g} \in \texttt{conv}(\mathscr{X})$. Figure 2(b) shows an example of natural neighbours for the point $\mathbf{g}$.

Natural neighbours inclusive (NNi): In some cases of non-uniformly distributed local areas, a training point which is far from the test point can be one of its natural neighbours, but a nearer point is excluded for its neighbourhood. To overcome this situation, natural neighbours inclusive has been proposed to include both the natural neighbours and those training points within the distance to the furthest natural neighbour. That is, $\mathscr{J}_{\mathbf{g}}^{NNi} = \{\mathbf{x}_j \in \mathscr{X} \,|\, \|\mathbf{g}-\mathbf{x}_j\| \leq \max_{\mathbf{x}_i \in \mathscr{J}_{\mathbf{g}}^{NN}} \|\mathbf{g}-\mathbf{x}_i\|\}$. Figure 2(d) is an example of natural neighbours inclusive.

IV. Monitoring Diesel fuels

In this section, we illustrate the effectiveness of Supervised Distance Preserving Projections coupled with Local Linear Regression based on neighbourhoods defined by k-nearest neighbours (LLR-kNN), enclosing k-nearest neighbours (LLR-ekNN), natural neighbours (LLR-NN) and natural neighbours inclusive (LLR-NNi). The SDPP is then compared with four state-of-the-art methods for unsupervised and supervised dimensionality reduction. For comparison, we considered Principal Component Analysis, Partial Least Squares, Kernel Principal Component Analysis and Kernel Partial Least Squares.

The application consists of estimating six different properties in summer diesel fuels starting from a set of spectral observations. The absorbance spectra were acquired by means of a spectrophotometer operating in the $900-1700$nm range, with a 2nm resolution. Each input observation consists of the 401-channel spectrum of absorbances ($\mathbf{x}_i \in \mathbb{R}^d$, with $d = 401$) and the corresponding outputs are the values of six different chemico-physical properties ($\mathbf{y}_i \in \mathbb{R}^m$, with $m = 6$): I) Boiling point; II) Cetane number; III) Density; IV) Freezing temperature; V) Total Aromatics; and, VI) Viscosity. The measurements of the product's properties were obtained in laboratory by reference methods. The dataset consists of $n = 135$ observations for learning the projection models and the local linear regression models and 125 observations for testing the results. The six outputs are modelled independently.

A. 2D projections and local linear regression

To get an insight on the spectra and their low-dimensional arrangement with respect to the six properties, we firstly projected the spectra ($d = 401$) onto a bi-dimensional subspace ($\mathbf{z}_i \in \mathbb{R}^s$, with $s = 2$) using PCA, KPCA, PLS, KPLS and the SDPP. Subsequently, the local linear regression methods were learned to regress the output responses onto the new low-dimensional input spaces. The parameters of both, the projection and the regression models, are estimated using training points only. 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 and, more importantly, to investigate the possibility to develop very parsimonious regression models afterwards. When kernel methods are used, Gaussian kernels are employed, with the 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 same heuristic is also used to define locality in LLR-kNN.

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 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 is obtained with unsupervised methods (PCA and KPCA), as per their stronger ability to arrange the projected input points on the basis of the response; an expected result. The bi-dimensional subspaces learned by the SDPP are based on two highly informative features that further emphasise this aspect of the projections. This is particularly true for the boiling point, 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, it seems that also for the SDPP, projections onto more features are needed.

The qualitative assessment of the projections can be quantified after recalling that when the dimensionality is reduced it is not necessarily possible to preserve all the similarities. From the point of view of LLR, the reduction causes a main kind of error: Data point that are not neighbours in $\mathscr{X}$ can be mapped close by in $\mathscr{Z}$, causing points to be falsely identified as similar. Such errors can be used to measure the trustworthiness of the $\mathscr{X} \rightarrow \mathscr{Z}$, which is defined by denoting with $U_{k_r}(i)$ the set of points that are in k_r-neighbourhood of $\mathbf{z}_i$ in $\mathscr{Z}$ but not in $\mathscr{X}$ and, with $r(i,j)$ the rank of $\mathbf{x}_j$ in the ordering based on its distance from $\mathbf{x}_i$. Trustworthiness of $\mathscr{X} \rightarrow \mathscr{Z}$ is then

$$M_{\text{trust}}^{\mathscr{X}\rightarrow\mathscr{Z}}(k_r) = 1 - C(k_r)\sum_{i=1}^{n}\sum_{j\in U_{k_r}(i)}(r(i,j)-k_r)$$

The neighbourhood size k_r is the amplitude of the region of interest, the term $C_(k_r)$ simply scales the measures in $[0,1]$. The upper row of plots in Figure 4 shows the trustworthiness of the projections for k_r ranging in the $[2,64]$ interval. The plots highlight how PCA and PLS are the best performers, with trustworthiness 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, their kernel extensions returned projections that are only moderately faithful. Similar results are also obtained by the SDPP, indicating that the apparent quality of the 2D displays does not imply a preservation of similarities between spectra.

This result is expected because such criterion is not in the SDPP's objective; the SDPP aims at mapping inputs characterised by similar outputs close by in the projection space. For regression it is, in fact, more desirable that the continuity of $\mathscr{Z} \rightarrow \mathscr{Y}$ is as high as possible. Such continuity can be defined by letting $V_{k_r}(i)$ be now the points that are in the k_r neighbourhood in $\mathscr{Z}$ but not in $\mathscr{Y}$ and, by letting $r(i,j)$

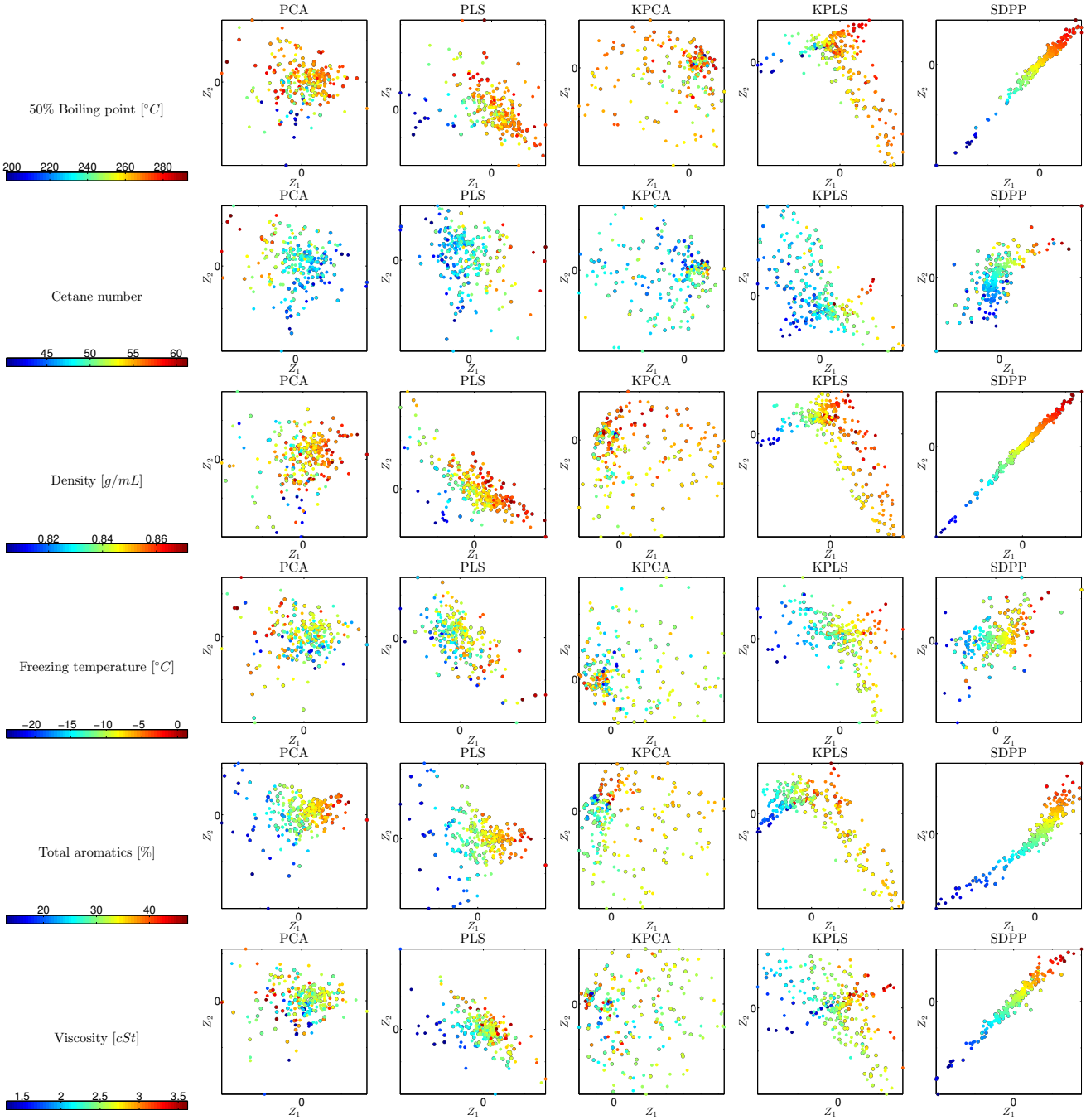


Fig. 3. Bi-dimensional projection and visualisation of the input spectra. Colouring based on output values is used to dye the inputs.

be the rank of y_j in the ordering based on its distance from y_i:

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

Note that here k_r and $C(k_r)$ bear the same meaning as before, whereas k is the locality parameter of the SDPP. The lower row of plots in Figure 4 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 the interval $[2,64]$ is used. The diagrams highlight how SDPP is consistently the best performer in representing the continuity between the projected spectra and their properties, 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 capture the input-output relationships.

The projection results suggest that, for all the responses, simple linear regression models calibrated either globally or locally over all the learning sample should be sufficient to

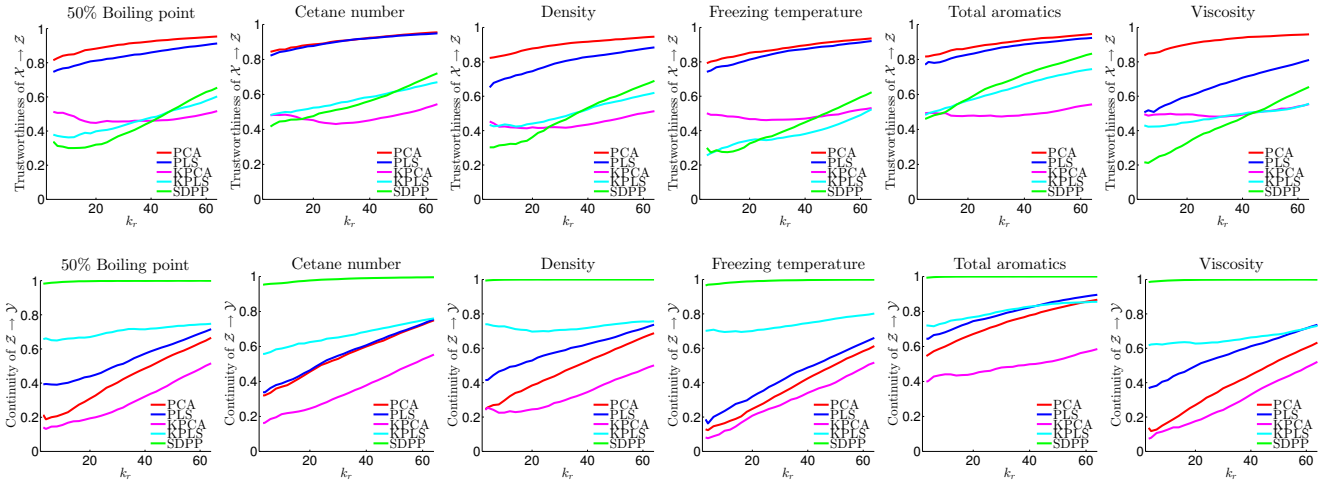


Fig. 4. Trustworthiness of the $\mathscr{X} \to \mathscr{Z}$ projection and continuity for the $\mathscr{Z} \to \mathscr{Y}$ regression, for a region of interest $k_r \in [2, 64]$.

estimate the fuel properties directly from the projected spectra.

TABLE I. TEST PERFORMANCE (RMSE).

Output	Models	SDPP	PLS	PCA	KPLS	KPCA
I	MLR	5.2e+0	1.0e+1	1.5e+1	1.3e+1	1.7e+1
	LLR-*k*NN	5.2e+0	1.1e+1	1.4e+1	1.3e+1	3.1e+1
	LLR-NN	5.2e+0	1.8e+1	1.7e+1	1.3e+1	3.8e+1
	LLR-e*k*NN	1.2e+1	1.8e+1	9.7e+0	1.3e+1	8.2e+1
	LLR-NN*i*	5.2e+0	9.7e+0	1.4e+1	1.3e+1	1.8e+1
II	MLR	4.0e+0	2.3e+0	2.3e+0	2.9e+0	3.5e+0
	LLR-*k*NN	3.8e+0	2.4e+0	2.4e+0	2.8e+0	4.9e+0
	LLR-NN	3.7e+0	2.6e+0	2.7e+0	2.8e+0	6.8e+0
	LLR-e*k*NN	3.8e+0	2.7e+0	2.9e+0	2.8e+0	4.0e+0
	LLR-NN*i*	3.8e+0	2.4e+0	2.5e+0	2.8e+0	3.8e+0
III	MLR	9.6e-4	5.1e-3	9.5e-3	7.1e-3	1.0e-2
	LLR-*k*NN	9.6e-4	5.2e-3	9.6e-3	7.3e-3	2.8e-2
	LLR-NN	9.6e-4	5.8e-3	9.5e-3	7.6e-3	1.7e-2
	LLR-e*k*NN	9.6e-4	1.1e-2	2.0e-2	7.5e-3	1.2e-2
	LLR-NN*i*	9.6e-4	5.3e-3	9.4e-3	7.4e-3	1.2e-2
IV	MLR	4.6e+0	3.6e+0	3.9e+0	3.4e+0	4.0e+0
	LLR-*k*NN	4.6e+0	4.1e+0	4.3e+0	3.5e+0	5.5e+0
	LLR-NN	4.6e+0	1.0e+1	5.1e+0	3.5e+0	5.2e+0
	LLR-e*k*NN	4.6e+0	4.8e+0	4.2e+0	3.5e+0	4.8e+0
	LLR-NN*i*	4.6e+0	4.0e+0	4.3e+0	3.5e+0	4.5e+0
V	MLR	7.9e-1	1.9e+0	2.3e+0	3.7e+0	5.7e+0
	LLR-*k*NN	7.7e-1	1.9e+0	2.8e+0	3.8e+0	8.2e+0
	LLR-NN	7.7e-1	2.1e+0	3.0e+0	3.7e+0	1.2e+1
	LLR-e*k*NN	7.8e-1	8.2e+0	2.3e+0	3.7e+0	8.2e+0
	LLR-NN*i*	7.7e-1	2.0e+0	2.6e+0	3.7e+0	6.0e+0
VI	MLR	1.4e-1	2.4e-1	3.9e-1	2.8e-1	3.9e-1
	LLR-*k*NN	1.4e-1	2.5e-1	4.1e-1	2.8e-1	1.6e+0
	LLR-NN	1.4e-1	2.6e-1	4.5e-1	2.8e-1	4.4e+0
	LLR-e*k*NN	1.4e-1	4.4e-1	2.6e-1	2.8e-1	7.1e-1
	LLR-NN*i*	1.4e-1	2.5e-1	4.2e-1	2.8e-1	5.5e-1

This was quantitatively verified after evaluating the accuracy of the MLR and the LLR-*k*NN, LLR-NN, LLR-e*k*NN and LLR-NN*i* models calibrated from all the obtained 2D-projections and, for each of the six outputs. The RMSE (Root Mean Square Error) results are reported in Table I. As expected, the SDPP nearly always leads to the smallest test errors, when both global MLR models and LLR models are considered. In that respect, it is important to notice that the accuracies achieved by MLR are already comparable with the accuracy of the analytical measurements. The improvements obtained by LLR although expected are thus very often negligible and of marginal importance in real-world applications.

V. CONCLUSIONS

The Supervised Distance Preserving Projection is a supervised dimensionality reduction method designed to project high-dimensional inputs onto a low-dimensional subspace where the geometry of the input points mimics the geometry of the output points. Such type of projection is desirable for designing parsimonious and yet accurate regression models from very high-dimensional and possibly correlated inputs in small sample problems, as those typically encountered in chemometrics. In this work, the applicability of the SDPP coupled with Local Linear Regression under these ill-posed regression conditions is investigated for a set of diesel fuels. On the basis of the experimental results, we found that the SDPP can generate informative and yet parsimonious projections finalised to the design of efficient calibration models.

REFERENCES

[1] J. J. J. Workman, "Review of process and non-invasive near-infrared and infrared spectroscopy: 1993-1999," *Appl. Spectrosc. Reviews*, vol. 34, pp. 1–89, 1999.

[2] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2002.

[3] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometr. Intell. Lab.*, vol. 58, no. 2, pp. 109–130, 2001.

[4] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[5] R. Rosipal and L. J. Trejo, "Kernel partial least squares regression in reproducing kernel Hilbert space," *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, 2002.

[6] Z. Zhu, T. Similä, and F. Corona, "Supervised distance preserving projections," *Neural Process. Lett.*, vol. In Press, 2013.

[7] C. J. Stone, "Consistent non-parametric regression," *Ann. Stat.*, vol. 80, pp. 595–645, 1977.

[8] M. R. Gupta, E. K. Garcia, and E. Chin, "Adaptive local linear regression with application to printer color management," *IEEE T. Image Process.*, vol. 17, pp. 936–945, 2008.

[9] Z. Zhu, F. Corona, A. Lendasse, R. Baratti, and J. A. Romagnoli, "Local linear regression for soft-sensor design with application to an industrial deethanizer," in *IFAC Proceedings Volumes*, 2011, pp. 2839–2844.