Regularization and improved interpretation of linear data mappings and adaptive distance measures

Marc Strickert
Computational Intelligence, Philipps Universität Marburg, DE

Barbara Hammer
CITEC centre of excellence, Bielefeld University, DE

Thomas Villmann
Comp. Intelligence Group, University of Applied Sciences Mittweida, DE

Michael Biehl
Intelligent Systems Group, University of Groningen, NL

***Abstract*—Linear data transformations are essential operations in many machine learning algorithms, helping to make such models more flexible or to emphasize certain data directions. In particular for high dimensional data sets linear transformations are not necessarily uniquely determined, though, and alternative parameterizations exist which do not change the mapping of the training data. Thus, regularization is required to make the model robust to noise and more interpretable for the user. In this contribution, we characterize the group of transformations which leave a linear mapping invariant for a given finite data set, and we discuss the consequences on the interpretability of the models. We propose an intuitive regularization mechanism to avoid problems in under-determined configurations, and we test the approach in two machine learning models.**

I. Introduction

Linear transformation of data can play a crucial role in machine learning algorithms. Principal component analysis (PCA) is a good example of a standard technique to preprocess data by a linear mapping before subsequent analysis, visualization or classification takes place [1], [2]. Alternative linear preprocessing schemes which aim at an extraction of semantically relevant contents of the data are addressed by popular techniques such as independent component analysis or slow feature analysis [3], [4]. These techniques are unsupervised in the sense that the linear data transformation is determined based on universal principles such as the reconstruction error of the given data set, statistical independence of the outputs, or slow variance over a temporal domain.

In particular in the case of supervised learning, linear data processing is often directly integrated into the target function and adaptation takes place based on the classification result. Examples range from simple feature selection schemes, which correspond to an axes parallel linear projection of the data [5], up to general linear mappings which transform the data space, possibly locally, such as in distance-based classifiers where the metric is adapted during training. Popular examples of the latter include metric learners such as generalized relevance learning vector quantization (GRLVQ), generalized matrix LVQ (GMLVQ), or metric adaptation schemes for k-nearest neighbor approaches [6]–[9].

In supervised scenarios, linear projections can be crucial due to several reasons: an additional linear transformation can enhance the capacity of the function class such that more flexible processing can be realized; this is the case in GRLVQ, for example [6]. Sensor and modern data acquisition technology often provide very high dimensional data. Popular examples are multi-spectral data, microarrays, document word matrices, etc. In these settings, the dimensionality of the data can easily exceed the number of available training samples, such that under-determined problems appear. Severe problems can occur for high dimensional data due to statistical effects, such as the empty-space phenomenon and similar [10]. A reasonable linear dimensionality reduction can help to avoid these problems, e.g. by preprocessing data with PCA, or as implicitly realized in several metric learners [11]. In these cases, linear preprocessing acts as a regularization technique to enhance the generalization ability of the technique and to deal with ill-posed learning scenarios.

Another, often very crucial aspect of linear transformations consists in its interface towards interpretability of the models. Restricted linear transformations such as axis parallel projection schemes or simple feature weighting directly offer rankings of input dimensions according to their relevance for a given task. Such ranking can directly be inspected by experts, who can verify whether this ranking corresponds to a semantically meaningful ordering of the given input features. This aspect constitutes a crucial part if a linear projection is used for feature selection algorithms such as proposed in [12], for example. Further, there exist very interesting applications, where features determined from a trained linear weighting scheme have a great potential as semantically meaningful markers, see e.g. [13]. An alternative is offered by linear transformations with low rank, since very powerful and intuitive discriminative data visualization techniques can be based thereon. These establish a direct visual interface to the learning results for the practitioner [14].

Interpretability of machine learning models becomes more and more important if complex learning scenarios are dealt with; in these settings, rather than a simple classification or regression task, an initially not precisely specified data analysis is carried out in which interesting aspects become apparent as a result of iterative/interactive data processing [15]–[17]. Here the human pattern recognition ability plays a crucial role in the data analysis loop, and intuitive interfaces to interpret the results are desirable endpoints for machine learning techniques. Even in more specific learning scenarios interpretability can play a crucial role to allow for an extraction of semantically

meaningful entities such as potential biomarkers in medical applications.

For a valid interpretation, it is essential that only reliable information is displayed to the practitioner, and contributions originating from noise or random effects are subtracted from the information. An interpretation which solely relies on random effects of the method can hardly lead to reliable information. Thus, uniqueness of the parts of the model which are interpreted by a human observer has to be guaranteed. First investigations in how far linear mappings can be interpreted have recently been proposed in [18]. Therein, Lagrange-based optimization is used to find an alternative mapping with a tradeoff between reconstruction quality and minimum norm, and relevance of mapping coefficients are assessed by sensitivity analysis, i.e. as contributions of coefficient derivatives to pairwise distance reconstruction.

In this contribution we address the question in how far mappings obtained in machine learning models lead to reliable information, thus, offer a valid interface to interpretable information. We will argue that this is not the case in general. We will focus on linear settings, where closed-form solutions can be provided, and we will show that, for a given finite data set, statistical invariances have to be factored out to guarantee this property. Rather than tightly integrating regularization schemes like ridge regression or Lasso [19] into a specific method we connect these invariances to classical principal component analysis, and we propose a simple method how to regularize a linear transformation during or after training at will. The benefit of this technique is investigated in two existing machine learning models, a linear regression model and a prototype-based classification scheme.

II. Linear mappings

We are interested in the uniqueness of linear data transformations which are embedded in any machine learning pipeline. In the general scenario, we consider a set of P feature vectors in N dimensions, concatenated in an $(N \times P)$-matrix X:

$$X = \left[\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^P\right] \quad \text{with} \quad \left\{\boldsymbol{x}^\mu \in \mathbb{R}^N\right\}_{\mu=1}^{P} . \quad (1)$$

A linear transformation of the data set can be parameterized in terms of a matrix $\Omega \in \mathbb{R}^{M\times N}$ with

$$\Omega\, \boldsymbol{x}^\mu = \boldsymbol{y}^\mu \in \mathbb{R}^M \quad \text{for all} \quad \mu = 1, 2, \ldots, P\,. \quad (2)$$

In vectorial notation: $\Omega\, X = Y$ with $Y = \left[\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^P\right] \in \mathbb{R}^{M\times P}$. Throughout the following we restrict the discussion to target dimensions $M \leq N$. We denote by $\boldsymbol{\omega}_k \in \mathbb{R}^N$ the rows of Ω, i.e. $\Omega = \left[\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_M\right]^\top$.

A mapping of the above form (2) naturally occurs in the framework of linear regression problems. Further, it occurs in machine learning models which include a linear transformation of the data such as metric-based approaches with adaptive metric [6]–[9], [20]–[23]. We will exemplarily consider a simple linear transformation and prototype-based models in the following. In these approaches, the learned matrix Ω is used as an indicator for the relevance of the input features: the overall contribution $\sum_i |\Omega_{iJ}|^2$ (or any other reasonable accumulator based on column J of matrix Ω) can serve as a measure of the relevance of feature J [7]. Further, for low-rank matrices, i.e. $M \in \{2, 3\}$, a direct visualization of the projections Y is possible and gives valuable insight into the relevant mutual ordering of the data [14].

The question now occurs whether this procedure leads to a valid interpretation; in other words: is the matrix Ω unique up to ambiguities which are contained in the learning task itself or is it subject to variances due to random effects only? Initial investigations of how to turn linear coefficients into interpretable results have been proposed in [18]. Here we rely on an alternative computationally simpler approach which focusses on invariance classes and their uniqueness of the considered linear transformations.

A. Column Space Projection

Obviously, a linear mapping and thus the matrix Ω is uniquely determined if and only if it is constrained on N base vectors, i.e. the rank of X is N. However, in general, we cannot assume this: on the one hand, situations where the data dimension exceeds the number of points become more and more common. On the other hand, typical features, for example adjacent channels in hyperspectral data sources, are not independent; thus, rather strong correlations are found and data are contained in a low-dimensional sub-manifold only. Hence, in general, the matrix Ω in (2) is not uniquely determined by X.

In the following we characterize invariances which yield the same mapping of given data X. To this end, we use a formalization which enables an intuitive regularization of Ω to a unique representation, and which allows an extension towards noise, i.e. cases where certain dimensions 'almost' vanish.

Assume matrices Ω and $\widetilde{\Omega}$ exist with $\Omega X = \widetilde{\Omega} X$, i.e. $D_\Omega X = 0$ where $D_\Omega = (\Omega - \widetilde{\Omega})$. Then,

$$D_\Omega X = 0 \leftrightarrow D_\Omega C D_\Omega^\top = 0 \leftrightarrow C(d_\Omega)_k = 0 \quad \forall k \quad (3)$$

with the symmetric positive semidefinite matrix $C = X\,X^\top \in \mathbb{R}^{N\times N}$ and $D_\Omega^T = [(d_\Omega)_1, \ldots, (d_\Omega)_M]$ denoting the rows of the difference matrix. Hence two matrices are equivalent if and only if all rows of the difference lie in the null space of C.

Without loss of generality, we can assume that all eigenvectors $\boldsymbol{u}_j$ of C are orthonormal and that the corresponding eigenvalues are ordered

$$\gamma_1 \geq \gamma_2 \ \ldots \geq \gamma_J > 0 = \gamma_{J+1} = \gamma_{J+2} = \ldots \gamma_N \quad (4)$$

Hence, C has J non-zero eigenvalues. In the case $N > P$, where the number of feature vectors is smaller than their dimension, the matrix C has at least $N - P$ zero eigenvalues. Also in data sets with $P > N$, correlated or interdependent features can result in a non-empty null space and $J < N$.

The eigenvectors $\mathbf{u}_j$ with eigenvalue zero, i.e. $j > J$, correspond to directions in feature space in which the data display no variation. Arbitrary linear combinations of these vectors can be added to the rows of Ω without changing the

projections Y in Eq. (2). Hence, a continuum of matrices Ω realizes the same mapping of the given data set.

Given a particular mapping $\Omega\, X \;=\; Y$ we can obtain a *reduced representation* $\widehat{\Omega}$ by means of projections into the column space of C. We define

$$\Psi \;=\; \left[\sum_{j=1}^{J} \boldsymbol{u}_j\, \boldsymbol{u}_j^{\top}\right] \;=\; \left[I - \sum_{j=J+1}^{N} \boldsymbol{u}_j\, \boldsymbol{u}_j^{\top}\right] \qquad (5)$$

with the $(N \times N)$ identity matrix I and eigenvectors $\boldsymbol{u}_j$ of C. The matrix $\widehat{\Omega}$ with

$$\widehat{\Omega} = \Omega\, \Psi, \quad \text{i.e.} \quad \widehat{\boldsymbol{\omega}}_k \;=\; \Psi\, \boldsymbol{\omega}_k \quad \text{for} \quad 1 \leq k \leq M, \qquad (6)$$

realizes the same mapping of the given data set, but contains no contributions from the null space of C.

It is straightforward to show that $\widehat{\Omega}$ from Eq. (6) is the solution of the following optimization problem:

$$\min \sum_{i=1}^{N}\sum_{j=1}^{M} \widetilde{\Omega}_{ij}^2 \quad \text{such that} \;\; \Omega\, X \;=\; \widetilde{\Omega}\, X. \qquad (7)$$

Obviously, rows $\widetilde{\boldsymbol{\omega}}_j$ of $\widetilde{\Omega}$ can be treated independently since $\sum_{ij} \widetilde{\Omega}_{ij}^2 \;=\; \sum_j \left|\widetilde{\boldsymbol{\omega}}_j\right|^2$. Now consider a vector $\widetilde{\boldsymbol{\omega}}_k$ with non-zero contributions from the null space of C, i.e. $\boldsymbol{u}_j^{\top}\widetilde{\boldsymbol{\omega}}_k \;=\; b_j \neq 0$ for at least one $j > J$. Exploiting the orthonormality of the $\boldsymbol{u}_j$ we observe for the projection $\widehat{\boldsymbol{\omega}}_k = \Psi\widetilde{\boldsymbol{\omega}}_k$

$$\left|\widehat{\boldsymbol{\omega}}_k\right|^2 \;=\; \left|\widetilde{\boldsymbol{\omega}}_k\right|^2 \;-\; \sum_{j=J+1}^{N} b_j^2 \;<\; \left|\widetilde{\boldsymbol{\omega}}_k\right|^2 .$$

This implies that the rows of the matrix Ω which solves (7) are orthogonal to the null space of C, since removing such contributions reduces the norm without changing the mapping. Hence $\widehat{\Omega}$ corresponds to the unique solution of problem (7).

Note that the formal solution $\widehat{\Omega}$ of the optimization problem (7) is well-known: Any mapping Ω on X can equivalently be characterized as $\widehat{\Omega} = X^{+}\, (\Omega\, X)$ with the Moore-Penrose pseudo-inverse X^{+} of the data matrix. The above, explicit formulation in terms of the projection Ψ facilitates the extension to regularization schemes which go beyond the exact elimination of null space contributions.

B. Regularization

The projection Ψ as introduced above eliminates all contributions from the null space of C. Note that the resulting mapping is identical for the given data X, but it puts a bias towards the most simple model regarding functionality on a separate training set for which invariance does not necessarily hold. In addition, as we will demonstrate below, an extension to the removal of small but non-zero eigenvalues can be advantageous under more general conditions.

Let us assume as above ordered eigenvalues such that γ_K is considered as small for some $K > J$. We denote the projection matrix $\Psi_K = \sum_{j=1}^{K} \boldsymbol{u}_j \boldsymbol{u}_j^{\top}$ and the corresponding regularization of a given matrix Ω as $\widehat{\Omega} = \Omega\, \Psi_K$. This projection retains the eigenspace corresponding to the K largest eigenvalues of C. It can be interpreted as a regularization of Ω which smoothens the mapping and the resulting projection $\widehat{\Omega}$ should be less specific to the particularities of the data set X. In analogy to (7), we obtain this matrix as unique solution of the optimization problem

$$\min \sum_{i=1}^{N}\sum_{j=1}^{M} \widetilde{\Omega}_{ij}^2 \quad \text{such that} \;\; \left\|\Omega\, X - \widetilde{\Omega}\, X\right\|_F^2 \leq \mathcal{L}(\Omega) \qquad (8)$$

where

$$\mathcal{L}(\Omega) = \sum_{i=1}^{M}\sum_{j=K+1}^{J} \lambda_j (\boldsymbol{u}_j^{\top}\boldsymbol{\omega}_i)^2 \qquad (9)$$

denotes the loss in the (squared) Frobenius norm when using the matrix where the space spanned by the vectors $\boldsymbol{u}_j$ for $j > K$ is projected out.

Note that

$$\widehat{\Omega}\, \boldsymbol{x} = (\Omega\Psi_K)\, \boldsymbol{x} = \Omega\, (\Psi_K \boldsymbol{x}) \;, \qquad (10)$$

which suggests two equivalent, complementary interpretations: either the regularized mapping is applied to the original feature vectors or the original mapping is applied to modified feature vectors. This means, either Ψ_K is used to regularize the linear mapping Ω while or after training. Alternatively, the feature vectors can be preprocessed using Ψ_K. Both interpretations are formally equivalent as regards the mapping, but the former offers the possibility to choose a suitable K while or after training, while this number has to be specified in advance if data are preprocessed. In addition, the numerical behavior of the algorithm can be different since a posterior regularization retains the high-dimensional search space; hence, possibly more ways are available to reach a minimum or local minimum along gradients of the cost function. This fact is not relevant in convex settings such as linear regression tasks, but it can play a role in more complex LVQ schemes, for example. Further, regularization of the linear mapping rather than the data opens the possibility to directly interpret the matrix entries as indicators of the relevance for the input features while a transformation of the features disrupts this possibility.

Note that this equivalence constitutes a bridge to the classical preprocessing technique PCA. If data are centered, a PCA projection of data exactly corresponds to the previous linear projection of data using the matrix Ψ_K.

C. Adaptive distance matrices

In recent years, a variety of techniques has been proposed to extend distance-based machine learning techniques by adaptive metrics, see e.g. [7]–[9], [20], [21], [24], [25]. In these settings, the metric is characterized by a general quadratic form, such that an equivalent formalization can be found in terms of a linear data transformation. This allows to transfer the above arguments immediately to the emerging field of distance learners.

As before, we consider a set of input vectors $X \;=\; [\boldsymbol{x}^1, \ldots, \boldsymbol{x}^P]$. These are possibly enriched by class labels if a supervised classification task is considered. For the moment,

the precise form of the machine learning problem is not relevant, only their potential for distance matrix regularization. We assume that the employed technique relies on pairwise distances of points, and we use the quadratic form

$$d(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^\top \Lambda (\boldsymbol{x} - \boldsymbol{y}) = [\Omega (\boldsymbol{x} - \boldsymbol{y})]^2 \qquad (11)$$

with positive semidefinite matrix $\Lambda \in \mathbb{R}^{N \times N}$ which can always be characterized as $\Lambda = \Omega^\top \Omega$ for some matrix $\Omega \in \mathbb{R}^{M \times N}$. The latter corresponds to a linear transformation of the data from $\mathbb{R}^N$ to $\mathbb{R}^M$ with $M \leq N$. Note that Ω is not uniquely determined by Λ, because usually several roots of Λ exist. This ambiguity is no problem in practice since a unique representation can be induced by the eigenvalue decomposition of Λ. This yields the unique symmetric positive semi-definite Ω with $\Lambda = \Omega^\top \Omega$ or a corresponding low-rank matrix where the space spanned by eigenvectors with eigenvalue 0 is disregarded, respectively.

Obviously, the above discussed concepts of column space projection and regularization apply immediately to parameterized distance measures of the form (11). Depending on the machine learning method, the overall function is determined by pairwise distances of data points $d(\boldsymbol{x}^i, \boldsymbol{x}^j)$ where $1 \leq i, j \leq P$ such as for k-nearest neighbor based approaches [8], [9], or the function relies on pairwise distances of data and prototypes $d(\boldsymbol{x}^i, \boldsymbol{w}^j)$ where $1 \leq i \leq P, 1 \leq j \leq k$ in prototype-based techniques. Here, $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^k$ constitute elements of $\mathbb{R}^N$ such as cluster centers for unsupervised vector quantization approaches [20], [21] or typical class representatives in supervised learning vector quantization [7], [24], [25]. For some Hebbian vector quantization update schemes, but not generally, the prototypes move in the linear space spanned by the given data points and the mixing coefficients can be restricted to convex combinations of the data [21].

Obviously, the functionality of such approaches is unchanged regarding training if the pairwise distances $d(\boldsymbol{x}^i, \boldsymbol{x}^j)$ and $d(\boldsymbol{x}^k, \boldsymbol{w}^l)$ are maintained. As before, an infinite number of equivalent matrices Ω and, consequently, Λ exists which provide the same functionality, namely all matrices which add contributions of the null space of C to rows $\boldsymbol{\omega}_j$ of Ω and the corresponding Λ.

In analogy to the regularization of a linear transformation, we can thus consider a regularized matrix

$$\widehat{\Omega} = \Psi \Omega \text{ and } \widehat{\Lambda} = \Omega^\top \Psi^\top \Psi \Omega \qquad (12)$$

with $\Psi = \sum_{j=1}^{J} \boldsymbol{u}_j \boldsymbol{u}_j^\top$ the projection to the non-vanishing eigendirections as before. This projection can be done while or after training, transforming the metric accordingly, or, alternatively, it can be applied priorly to the given data set. The latter also implicitly projects prototypes since its adaptation are restricted to an according subspace of the data space.

As before, we can use this observation as a motivation to regularize the model by dividing out Ψ_K instead of Ψ, referring to the smallest K (possibly nonzero) eigenvalues and eigenvectors of C. Again, this leads to a formalization which is not identical to the original setting but, hopefully, widely equivalent, dividing out invariances and noise in the data.

In the latter setting, an equivalence of prior and posterior projection of data is lost since the nonzero components can influence the dynamics of the learning algorithm. In addition, the effect of this regularization on the prototypes can be enhanced as compared to its influence on the data: for a prototype of the form $\boldsymbol{w}^j = \sum_i \alpha_{ji} \boldsymbol{x}^i$ we find

$$\|\boldsymbol{w}^j - \Psi_k \boldsymbol{w}^j\|^2 = \sum_{l=1}^{P} \alpha_{jl}^2 \sum_{j=K+1}^{J} (\boldsymbol{u}_j^\top \boldsymbol{x}^l)^2 . \qquad (13)$$

Hence, if the scaling term α_{jl}^2 is large for some $K+1 \leq l \leq J$ or, in other words, the direction $\boldsymbol{u}_l$ is prominent for the prototype $\boldsymbol{w}^j$, the loss in accuracy can be significant. Because of this observation, it might be advisable to base the regularization by means of Ψ_k on the vectors $[\boldsymbol{x}^1, \ldots, \boldsymbol{x}^P, \boldsymbol{w}^1, \ldots, \boldsymbol{w}^k]$ and corresponding C rather than the data alone. The same would apply to training schemes where prototypes cannot be represented or approximated as linear combinations of the feature vectors. In scenarios where prototypes $\boldsymbol{w}^j$ change their position, only regularization during training or posterior regularization is possible.

D. Interpretation of Linear Mappings and Relevance Matrices

Frequently, the matrix Λ, cf. Eq. (11), or the linear transformation Ω, cf. Eq. (2) are interpreted as 'significance' of single features and feature pairs. Common schemes involve $\sum_i \Omega_{is}^2$ with row index s in case of a linear transformation. For a symmetric positive semidefinite representation matrix Λ of Ω this corresponds to diagonal elements $\Lambda_{ss} = \sum_i \Omega_{is}^2$ of an adaptive distance measure. Values Λ_{ss} can be interpreted as a measure of the relevance of feature s for the classification. This has been successfully employed for selecting relevant biomarkers in the medical context, see for example [13], [26].

Clearly, contributions from the null space of C can yield misleading results in this respect. Consider, as simple extreme example, two features which are perfectly correlated in the data set, i.e. $x_s^\mu = x_t^\mu$ for all $\mu = 1, 2 \ldots P$. Assume furthermore that these features have no discriminatory value at all, their values could be random numbers independent of the class membership, for instance. Any matrix Δ which satisfies $\Delta_{ks} = -\Delta_{kt}$ can be added to Ω without changing the linear mapping, pairwise distances, or the classification of training examples. However the magnitudes of $\Lambda_{ss} = \sum_k \Omega_{ks}^2$ and Λ_{tt} vary explicitly with the value of Δ_{ks} and Δ_{kt}. A naive analysis of Λ could, therefore, assign high discriminative power to completely irrelevant features.

This pitfall is clearly avoided when applying the column space projection (6) with matrix C after or during training. In this case, taking into account the scaling of data as concerns feature dimension s, the value $\sum_i \Omega_{is}^2$ corresponding to the entry Λ_{ss} or any other reasonable accumulator of column s of the transformation matrix Ω takes into account the overall contribution of feature s to the overall mapping which cannot be explained by purely statistical effects.

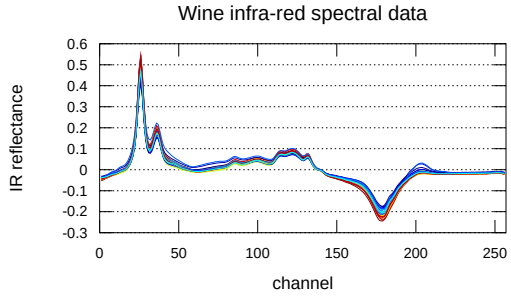

Fig. 1. 256-dimensional data of near-infrared wine spectra. Color corresponds to alcohol content to be predicted: blue means low alcohol content, red refers to high alcohol concentrations.

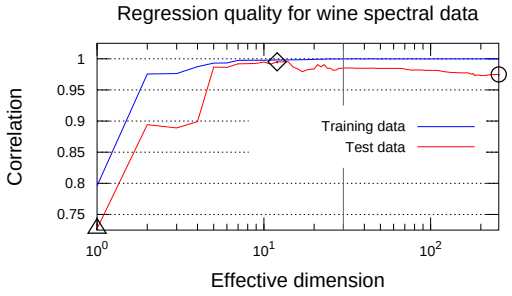

Fig. 2. Regression quality for training data and test data, expressed as correlation of prediction and ground truth. The triangle highlights over-regularization of the linear mapping coefficients to a singe effective dimension, the diamond indicates best test set correlation for 12 effective dimensions, and the circle corresponds to non-regularized coefficients. The vertical line marks the transition of effective dimensionalities below and above the number of training samples.

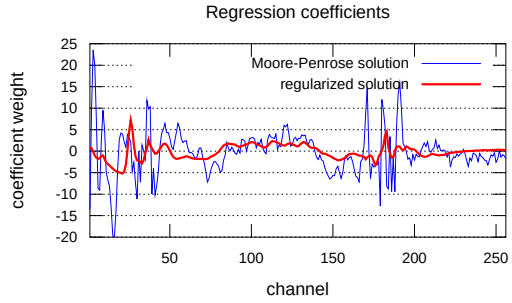

Fig. 3. Comparison of regression coefficients based on Moore-Penrose pseudo-inverse and the coefficient vector projected onto an effective data dimensionality of 12.

III. Example Applications

We demonstrate the effect of regularization in two popular machine learning models: a linear regression, and a generalized matrix LVQ (GMLVQ) classifier which relies on an adaptive metric of the data. Regularization is studied for real-world problems of under-determined data sets with less samples than attributes. One case corresponds to hyper-spectral data with highly correlated adjacent frequency channels. The other case investigates effects in physico-chemical property prediction of molecular data. For classification, percentile binning of the regression targets into two and three classes is carried out. Prior to method application, data centers of the training sets are determined and subtracted from training and test sets to ensure valid eigen-decompositions of centered data covariance matrices.

A. Regularization of Linear Mappings

For linear regression problems $\Omega\, X = Y$ with few observations composed of many attributes, under-determined systems of linear equations have to be solved. The Moore-Penrose pseudo-inverse X^+ of the data matrix generally yields robust solutions for a valid mapping matrix Ω:

$$\Omega\, X = Y \quad \rightarrow \quad \Omega = X^+ Y\,. \tag{14}$$

The obtained matrix Ω is known to provide minimum Frobenius norm like required in (7). Still, posterior regularization acts beneficially on generalization properties as shown in the following for exemplary regression of near-infrared spectral data and a biochemical regression problem on human intestinal absorption based on physico-chemical compound descriptors.

1) Near-infrared spectral data: Figure 1 contains 256-dimensional spectra of 124 wine samples [27], split into 94 training and 30 test samples. In agreement with [28] samples number 34, 35, and 84 where discarded as outliers from the training set. Finally, the roles of training and test sets are switched for enhancing the under-determination problem in the examples. The regression task is to predict their alcohol contents being assigned to the spectral profiles as color shades in Figure 1. A test set of 91 unseen samples is used for assessing generalization performance.

Due to under-determination, there is an infinite number of perfect solutions to the regression problem for the training data set, and further regularization is applied to the solution based on the Moore-Penrose inverse. As shown in Figure 2 the training performance regularized to given effective dimensions is excellent for more than one effective dimension, and test set correlations are also very good, i.e. greater than 0.95, for effective dimensionalities of five or higher. The term 'effective dimension' is used in regularization to refer to the controlled rank of the data matrix with a natural rank J corresponding to the number of positive eigenvalues. An optimum effective dimensionality of 12 can be identified (diamond) for the test set between the extremes of a over-regularization with a single effective dimension (triangle) and 256 dimensions, i.e. no regularization (circle). The slight change of test performance beyond the right vertical line indicates the influence of the null-space contributions on the test set while the training set performance is not affected.

Given regression weights for the optimum effective dimensionality of 12 a comparison with the original regression weights is shown in Figure 3. Regularization yields a much smoother weight vector in which some of the strongly peaked original channel weights are completely cancelled. As a consequence, its squared Euclidean norm is only 758.0 in contrast to 7132.8 for the original weights. Since the output is only one-dimensional, i.e. $M = 1$, the relevance of input dimension s can be judged by the absolute value of the linear weight as displayed in Figure 3. Obviously, the original

Moore-Penrose pseudo-inverse solution still leads to a large number of possibly relevant dimensions, albeit the null space of data is already divided out for this setting. Using further regularization, a much clearer profile with less candidates s results where, in particular, local fluctuations are dampened.

2) QSAR prediction: Human intestinal absorption (HIA) is an important prediction problem in quantitative structure-activity relationship (QSAR) modeling. A data set with 127 chemical compounds and 1499 molecular descriptors per compound is taken from [29]. In order to compensate for the very different descriptor value domains and density distributions, each attribute vector was replaced by ordinal ranks of the observed values, thereby removing physical dimensions like mass, charge, or volume. Despite the loss of meaningful molecular information, this pre-processing turned out to provide comparable linear attribute weights. For modeling regression, the data set is randomly split into 75% of training samples and 25% of test samples.

Figure 4 shows a strong dependence of the regression quality of unseen data on the effective dimension. Everything related to the null-space, that is for effective dimensionalities greater than the training set size 114, rather constant correlations (top panel) and mean square errors (MSE, bottom panel) are observed for training and test sets. While a monotonic decrease of the correlation is seen for the training set for smaller effective dimensionalities, a massive improvement of the test set performance is found for intermediate ranges. Maximum test set correlation of 0.792 is obtained at 56 effective dimensions; a minimum MSE of 0.374, which is only about a fifth amount of the non-regularized coefficients with MSE of approximately 2, is achieved at 17 effective dimensions. For the latter number, the ratio of the squared Euclidean norms of the optimum regularized mapping coefficients Ω_r over the original coefficients Ω is decreased by a factor of about 42. The corresponding weight vectors (not shown here due to lack of space) display a behavior which is qualitatively very similar to the relevance profile obtained for the near-infrared spectral data: a reduction of dimensionalities caused by small eigenvalues results in a strong decrease of the number of pronounced peaks of the relevance profile, thus indicating that these weightings are attributed to noise in the data rather than to functional causality.

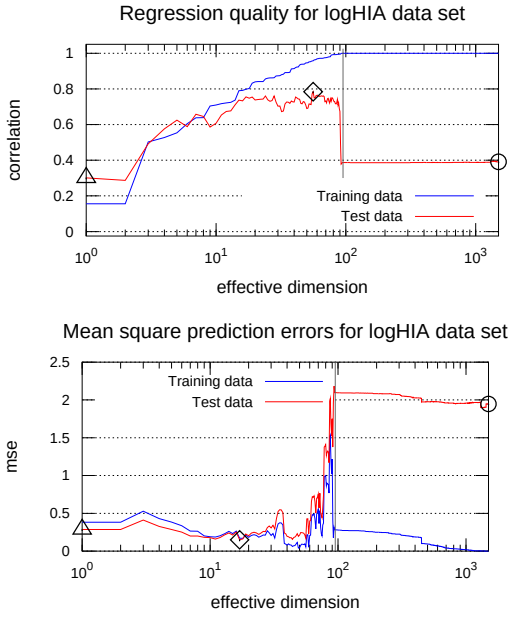

Fig. 4. Regression quality for training data and test data of human intestinal absorption. The top panel shows correlation and the bottom panel refers to MSE of regularized mappings depending on the effective dimension. The diamond indicates highest test set correlation (top panel) for 56 effective dimensions and lowest MSE (bottom panel) for 17 effective dimensions. Circles correspond to non-regularized coefficients, the triangle to a single effective dimension. Vertical lines mark the transition of effective dimensionalities below and above the number of training samples.

B. Regularization of discriminative distance measures

To demonstrate the effect of regularization on adaptive distance measures, we exemplarily address Generalized Matrix LVQ (GMLVQ) as introduced in [7]. The method deals with a classification task, i.e. data $\boldsymbol{x}^i$ with assigned class labels $y_i \in \{1, \ldots, C\}$. A GMLVQ model is characterized by a finite number of prototypes $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^k \in \mathbb{R}^N$ enriched with class labels $c(\boldsymbol{w}^i) \in \{1, \ldots, C\}$. In addition, an adaptive matrix $\Lambda = \Omega^\top \Omega$ parameterizes the distance

$$d_\lambda(\boldsymbol{x}, \boldsymbol{w}) = (\boldsymbol{x} - \boldsymbol{w})^\top \Lambda (\boldsymbol{x} - \boldsymbol{w}) . \tag{15}$$

A GMLVQ model assigns a given data point to the class of its nearest prototype as measured by this metric

$$\boldsymbol{x} \mapsto c(\boldsymbol{w}^i) \text{ such that } d_\Lambda(\boldsymbol{x}, \boldsymbol{w}^i) \text{ is minimal.} \tag{16}$$

Training takes place by an optimization of the cost function

$$E = \sum_{i=1}^{P} \Theta \left(\frac{d_\Lambda(\boldsymbol{x}, \boldsymbol{w}^+) - d_\Lambda(\boldsymbol{x}, \boldsymbol{w}^-)}{d_\Lambda(\boldsymbol{x}, \boldsymbol{w}^+) + d_\Lambda(\boldsymbol{x}, \boldsymbol{w}^-)} \right) \tag{17}$$

where $\boldsymbol{w}^+$ ($\boldsymbol{w}^-$) denotes the closest prototype with the same label as $\boldsymbol{x}^i$ (a different label than $\boldsymbol{x}^i$). Depending on the discrimination parameter σ in the squashing function $\Theta(x) = \frac{1}{1+\exp(-\sigma \cdot x)} \in [0, 1]$ the cost function approximates classification error counts corresponding to the setting $d_\Lambda(\boldsymbol{x}, \boldsymbol{w}^+) < d_\Lambda(\boldsymbol{x}, \boldsymbol{w}^-)$ and, in addition, maximizes the hypothesis margin as characterized by the numerators of the summands. Training takes place by means of a gradient technique with respect to both prototype locations and metric parameters Λ. For simplicity, we regularize only with respect to the matrix $X X^\top$, cf. Eq. 1, neglecting the potential influence of prototypes outside the span of feature vectors. The GMLVQ classifier available at https://mloss.org/software/view/323/ is used to optimize discriminative class prototypes and the matrix metric Λ. One prototype per class is chosen, and the discrimination parameter in GMLVQ is set to $\sigma = 50$.

1) Near-infrared spectral data: Alcohol levels in the regression problem are converted into three bins of equal size to be considered as classification targets. Despite of binning,

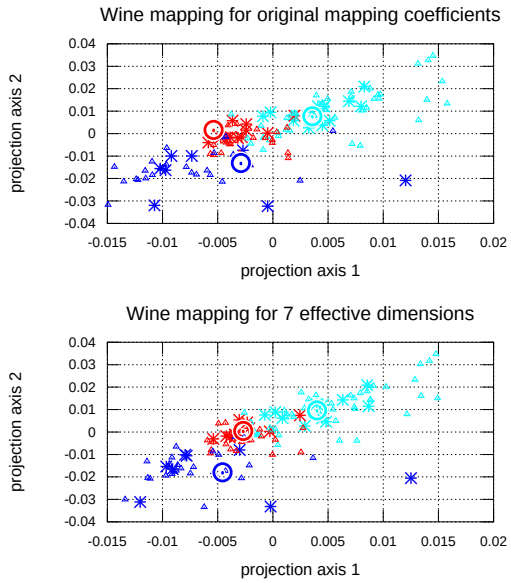

Fig. 5. Projection of wine spectrum data to original (top) and regularized (bottom) subspace Ω. Colors refer to the three classes of wine alcohol content. Asterisks (*) refer to training samples, triangles ($\triangle$) refer to test samples, and circles ($\odot$) are the class prototypes.

no natural semantic meaning of low, medium, and high alcohol content is established, because GMLVQ treats different labels as indicators of independent data subsets. Such an independence assumption complicates the classification problem, because misclassification of adjacent bins is equally rated to confusion of low and high content. Thus, the classification problem is used for illustration only. A subspace of $M = 2$ dimensions is considered, allowing GMLVQ for adopting a (2×256) parameter matrix Ω, i.e. a rank-2 matrix metric Λ.

A typical subspace resulting from GMLVQ is shown in the top panel of Figure 5. The panel below contains a regularized subspace mapped by Ω_r using an effective data dimension of 7. It can be seen that the prototypes in the regularized subspace are more aligned along a line, thus, emphasizing the natural correlation of data vectors and alcohol content. The choice of effective dimension was taken by looking at Figure 6: as expected, the training classification error increases for smaller effective dimensionalities as effect of regularization; at the same time, a minimum test error is found for the rank of the training data matrix limited to 7.

A dramatic effect of regularization on the mapping coefficients Ω and, consequently, on the matrix metric Λ is shown in Figure 7. Without regularization (top left panel), attribute pairs exhibit a strong checkerboard pattern while the much clearer and spatially consistent picture for regularization (top right panel) indicates interesting attributes around channel indices 25 and 180. Accumulating pairwise attribute interactions to the diagonal to characterize the linear mapping, as discussed above, smoothness and smaller coefficients $\Lambda_{r,ss}$ are obtained by regularization (bottom panel).

2) QSAR data: The data set on human intestinal absorption was randomly split into 75% training and 25% test samples,

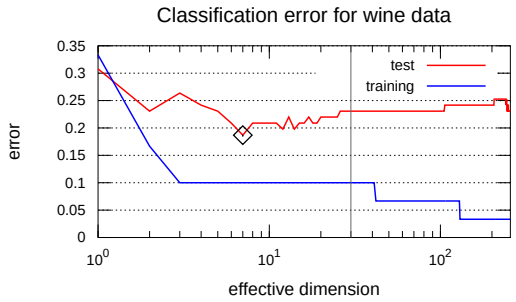

Fig. 6. Classification error for training and test set depending on coefficient regularization. The vertical lines marks the rank of the training data matrix.

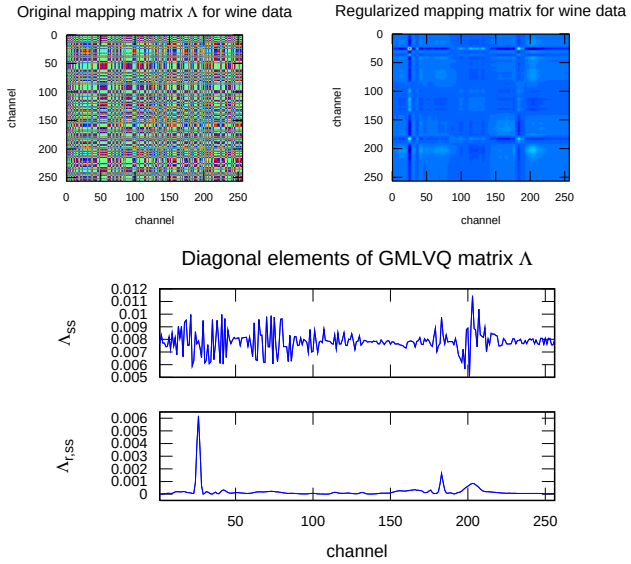

Fig. 7. Wine data mapping coefficients for classification of 256-dimensional spectra. Top: original matrix metric $\Lambda = \Omega\Omega^\top$ (left) and its regularized counterpart Λ_r for an effective dimension of 7. Red color represents large entries, dark blue represents small values. Bottom: corresponding diagonal matrix elements, i.e. indication of classification-specific attributes.

and the original regression target (HIA) was binned into two classes of equal size. Due to its inherent complexity and large dimensionality, the data split affects the classifier training, and representative results are chosen for presentation in Figure 8. As can be seen, after training and without regularization the

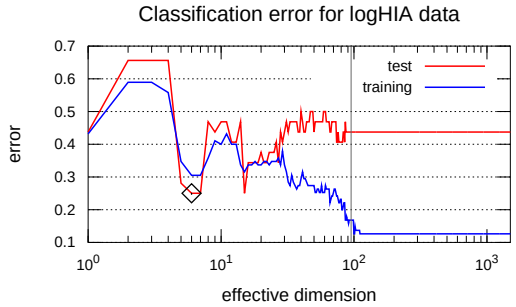

Fig. 8. Classification error for training data and test data of human intestinal absorption. The diamond indicates lowest test set error for 6 effective dimensions.

test error is at about 43%, i.e. just below guessing. At the same time, training performance at a 12% error rate is quite good. By reducing the effective dimensionality below the natural data rank the training error increases, while the test error decreases to 25% at an effective dimensionality of 6. A nice thing to be observed is the consistently coupled change of training and test set quality until an effective dimension of 25; this shows how the effective dimension acts as globally valid control parameter on the classification model.

IV. Discussion and Summary

In this paper we have addressed the uniqueness of linear mappings embedded in machine learning algorithms and its consequences on the interpretability of the mapping coefficients. We have characterizes equivalence classes of linear transformations on finite data sets which can be very rich in particular if high dimensional data or an intrinsically low dimensional data manifold with correlations of the features are dealt with. We have proposed a computationally efficient regularization scheme which avoids this problem, and demonstrated the feasibility of the approach in two machine learning scenarios and two benchmark data sets.

The regularization experiments illustrate how posterior regularization may enhance generalization performance of solvers for under-determined systems although they are known to be optimum on the training set. In common machine learning scenarios, the best effective dimensionality might be assessed for a given quality criterion based on a separate validation set, if available. Thus, the proposed procedure opens the way towards an automated process.

The experiments focused on posterior regularization. In GMLVQ classification, regularization during training based on both training data and prototypes is an important topic of future research, because a substantial change of convergence dynamic may occur depending on the data complexity.

Finally, the approach seems suited not only for the scenario of a global linear mapping. Local linear approaches such as proposed in the context of local metric learning could be addressed in a similar way.

Acknowledgement

BH gratefully acknowledges funding by DFG under grants number HA2719/6-1 and 7-1 and by the CITEC centre of excellence. MS is kindly supported by the LOEWE excellence program SYNMIKRO.

References

[1] J. Wu and J. Wang, "PCA-based SVM for automatic recognition of gait patterns," *J Appl Biomech*, vol. 24, no. 1, pp. 83–7, 2008.

[2] L. van der Maaten, E. Postma, and H. van den Herik, "Dimensionality reduction: A comparative review," Tilburg University Technical Report, TiCC-TR 2009-005, Tech. Rep., 2009.

[3] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, no. 4-5, pp. 411–430, 2000.

[4] L. Wiskott and T. Sejnowski, "Slow Feature Analysis: Unsupervised Learning of Invariances." *Neural Computation*, vol. 14, no. 4, pp. 715–770, 2002.

[5] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[6] B. Hammer and T. Villmann, "Generalized relevance learning vector quantization," *Neural Networks*, vol. 15, no. 8-9, pp. 1059–1068, 2002.

[7] P. Schneider, M. Biehl, and B. Hammer, "Adaptive relevance matrices in learning vector quantization," *Neural Computation*, vol. 21, pp. 3532–3561, 2009.

[8] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[9] K. Weinberger, F. Sha, and L. Saul, "Convex optimizations for distance metric learning and pattern classification," *Signal Processing Magazine*, vol. 27, pp. 146–158, 2010.

[10] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer, 2007.

[11] M. Biehl, K. Bunte, F.-M. Schleif, P. Schneider, and T. Villmann, "Large margin discriminative visualization by matrix relevance learning," in *Proc. IEEE World Congress on Computational Intelligence, WCCI 2012*.

[12] D. Hardin, I. Tsamardinos, and C. F. Aliferis, "A theoretical characterization of linear SVM-based feature selection," in *Proceedings of the twenty-first international conference on Machine learning*, ser. ICML '04. New York, NY, USA: ACM, 2004, pp. 48–55.

[13] W. Arlt, M. Biehl, A. Taylor, S. Hahner, R. Libe, B. Hughes, P. Schneider, D. Smith, H. Stiekema, N. Krone, E. Porfiri, G. Opocher, J. Bertherat, F. Mantero, B. Allolio, M. Terzolo, P. Nightingale, C. Shackleton, X. Bertagna, M. Fassnacht, and P. Stewart, "Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors," *J Clinical Endocrinology and Metabolism*, vol. 96, pp. 3775–3784, 2011.

[14] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl, "Limited rank matrix learning, discriminative dimension reduction, and visualization," *Neural Networks*, vol. 26, pp. 159–173, 2012.

[15] M. Ward, G. Grinstein, and D. A. Keim, *Interactive Data Visualization: Foundations, Techniques, and Application*. A. K. Peters, Ltd, 2010.

[16] A. Vellido, J. Martin-Guerroro, and P. Lisboa, "Making machine learning models interpretable," in *ESANN'12*, 2012.

[17] S. Rüping, "Learning interpretable models," Ph.D. dissertation, Dortmund University, 2006.

[18] M. Strickert and M. Seifert, "Posterior regularization and attribute assessment of under-determined linear mappings," in *European Symposium on Artificial Neural Networks (ESANN)*, M. Verleysen, Ed. D-facto Publications, 2012, pp. 67–72.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, pp. 267–288, 1996.

[20] A. Gisbrecht and B. Hammer, "Relevance learning in generative topographic mapping," *Neurocomputing*, vol. 74, no. 9, pp. 1359–1371, 2011.

[21] B. Arnonkijpanich, A. Hasenfuss, and B. Hammer, "Local matrix adaptation in topographic neural maps," *Neurocomputing*, vol. 74, no. 4, pp. 522–539, 2011.

[22] J. Ye, Z. Zhao, and H. Liu, "Adaptive distance metric learning for clustering," in *IEEE Conf. on Computer Vision and Pattern Recog.*, 2007.

[23] R. Jin, S. Wang, and Y. Zhou, "Regularized distance metric learning: Theory and algorithm," in *Advances in Neural Information Processing Systems 22*, 2009.

[24] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl, "Regularization in matrix relevance learning," *IEEE Transactions on Neural Networks*, vol. 21, pp. 831–840, 2010.

[25] P. Schneider, M. Biehl, and B. Hammer, "Distance learning in discriminative vector quantization," *Neural Computation*, vol. 21, pp. 2942–2969, 2009.

[26] M. Biehl, P. Schneider, D. Smith, H. Stiekema, A. Taylor, B. Hughes, C. Shackleton, P. Stewart, and W. Arlt, "Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors," in *20th European Symposium on Artificial Neural Networks (ESANN 2012)*, M. Verleysen, Ed. d-side publishing, 2012, pp. 423–428.

[27] UCL, "Spectral Wine Database," Provided by Prof. Marc Meurens, Université Catholique de Louvain, http://www.ucl.ac.be/mlg/, 2007.

[28] C. Krier, D. François, F. Rossi, and M. Verleysen, "Feature clustering and mutual information for the selection of variables in spectral data," in *European Symposium on Artificial Neural Networks (ESANN)*. D-side Publications, 2007, pp. 157–162.

[29] A. Soto, R. Cecchini, G. Vazquez, and I. Ponzoni, "Multi-Objective Feature Selection in QSAR Using a Machine Learning Approach," *QSAR & Combinatorial Science*, vol. 28, no. 11–12, pp. 1509–1523, 2009.