ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Posterior regularization and attribute assessment of under-determined linear mappings

Marc Strickert[a] and Michael Seifert[b]

[a] Knowledge Engineering and Bioinformatics, University of Marburg, DE
[b] Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben, DE

Abstract. Linear mappings are omnipresent in data processing analysis ranging from regression to distance metric learning. The interpretation of coefficients from under-determined mappings raises an unexpected challenge when the original modeling goal does not impose regularization. Therefore, a general posterior regularization strategy is presented for inducing unique results, and additional sensitivity analysis enables attribute assessment for facilitating model interpretation. An application to infrared spectra reflects data smoothness and indicates improved generalization.

1 Introduction

While the machine learning community heads for scientific contributions referring to the power of non-linear models, the engineering community, for example, tends to prefer the simplicity and interpretability of linear or linearized models. Certainly, linear mappings allow the structurally simplest way to consider each attribute in a transformation of vector data. Several standard tools like principal component analysis, independent component analysis, linear discriminant analysis, or partial least squares regression show that linear approaches have a respectable modeling power [1]. However, whole books on factor analysis [4] indicate that the interpretation of mapping coefficients is often an intriguing challenge. For example, rotational transformation contributions of coefficients do not influence to, but obfuscate, the goal of separating class projections.

During the last decade, linear distance metric learning has become an area of active research. In neighborhood component analysis and large margin nearest neighbor classification label information is used to tune linear mappings for enhancing class discrimination of transformed data [3]. Learning data prototypes and their metric simultaneously for given data is realized in matrix learning vector quantization [8], and correlative matrix mapping seeks linear subspaces for optimum multivariate regression between distance spaces [9]. Also, unsupervised adaptive distance metric learning exist for optimizing clusterability [12]. Only few approaches provide norm-reducing parameter regularization [5], because of the conflict between minimum null vectors and highly structured solutions.

Each above-mentioned model is driven by linear mapping operations, being implicitly involved in cases of matrix metric learners. In the following we refer to sources of linear mappings as external models. Using only coefficient matrices and source data, the aim of this work is to provide a better understanding of the given mappings by applying regularization, standardization, and assessment of the data attributes independently of the original mapping methods.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2 Methods

Three steps are proposed to analyze a given mapping. The first step aims at a reconstruction of the given linear mapping by a regularized one. This can be interpreted as posterior application of Tikhonov regularization (ridge regression) and is related to the least absolute shrinkage and selection operator (LASSO) [10]. The second step of rotation standardization is recommended in cases where rotation does not play a role, such as for classification of projected samples. The final step involves attribute assessment for posterior analysis of linear mappings.

2.1 Posterior regularization of under-determined linear mappings

Let N input vectors be given as rows $\mathbf{x}^j \in \mathbf{X} \subset \mathbb{R}^M, \mathbf{x}^j = (x^j_k)_{1\le k\le M}, 1 \le j \le N$, and let $\boldsymbol{\omega} \in \mathbb{R}^{M\times u}$ be a matrix with M rows corresponding to the number of data attributes, and with u columns $\boldsymbol{\omega}^k$, $1 \le k \le M$, corresponding to the dimension of the linear mapping defined by $\boldsymbol{p} = \mathbf{X}\cdot\boldsymbol{\omega}$. For fewer sample vectors than data attributes, i.e. for $N < M$, this under-determined system with target $\boldsymbol{p}$ allows many different, yet, equivalent mapping solutions for $\boldsymbol{\omega}$.

If parameters $\boldsymbol{\omega}$ are free to be positive or negative, a unique matrix $\boldsymbol{\omega}^*$ is sought with $\boldsymbol{p} = \mathbf{X}\cdot\boldsymbol{\omega}^* = \mathbf{X}\cdot\boldsymbol{\omega}$ such that its squared Frobenius norm $\|\boldsymbol{\omega}^*\|^2_F$ is minimum. A formulation of this constraint optimization problem with Lagrange multiplier λ is provided separately for each column $\boldsymbol{v} = \boldsymbol{\omega}^k$ as [2]

$$\mathscr{L}(\boldsymbol{v}^*, \lambda) = \sum_{l=1}^{M} \left(\boldsymbol{v}^*_l\right)^2 + \lambda\cdot\alpha\cdot\sum_{j=1}^{N}\left(\mathbf{x}^j\cdot(\boldsymbol{v}-\boldsymbol{v}^*)\right)^2. \quad (1)$$

Contrary to LASSO shrinkage, putting the Lagrange constraint on the mapping reconstruction rather than on the norm of the coefficient vector allows expanding an initial null vector $\boldsymbol{v}^*$ until the mapping constraints get fulfilled. The constraint weight α helps during optimization for compensating very different contributions of norm and constraint terms. Saddle points of $\mathscr{L}$ are found by unconstrained minimization of the squared norm of its $(M+1)$-dimensional gradient $\boldsymbol{L} = (\partial\mathscr{L}/\partial\boldsymbol{v}^*, \partial\mathscr{L}/\partial\lambda)$:

$$\boldsymbol{L}(\boldsymbol{v}^*, \lambda) = \left(2\cdot\boldsymbol{v}^* - 2\cdot\lambda\cdot\alpha\cdot\mathbf{X}^{\mathsf{T}}\cdot\left(\mathbf{X}\cdot(\boldsymbol{v}-\boldsymbol{v}^*)\right),\ \alpha\cdot\sum_{j=1}^{N}\left(\mathbf{x}^j\cdot(\boldsymbol{v}-\boldsymbol{v}^*)\right)^2\right). \quad (2)$$

The partial derivatives of $G(\boldsymbol{v}^*, \lambda) = |\boldsymbol{L}|^2_2$ for gradient-based optimization are:

$$\frac{\partial G(\boldsymbol{v}^*,\lambda)}{\partial \boldsymbol{v}^*} = 4\cdot\left(\frac{\partial\mathscr{L}}{\partial\boldsymbol{v}^*} + \lambda\cdot\alpha\cdot\mathbf{X}^{\mathsf{T}}\cdot\left(\mathbf{X}\cdot\frac{\partial\mathscr{L}}{\partial\boldsymbol{v}^*}\right) - \frac{\partial\mathscr{L}}{\partial\lambda}\cdot\boldsymbol{U}\right), \quad (3)$$

$$\frac{\partial G(\boldsymbol{v}^*,\lambda)}{\partial \lambda} = -4\cdot\boldsymbol{U}^{\mathsf{T}}\cdot\frac{\partial\mathscr{L}}{\partial\boldsymbol{v}^*} \quad \text{with} \quad \boldsymbol{U} = \alpha\cdot\mathbf{X}^{\mathsf{T}}\cdot\left(\mathbf{X}\cdot(\boldsymbol{v}-\boldsymbol{v}^*)\right). \quad (4)$$

Formally dispensable parentheses are added for improved computational efficiency. For optimization the quasi Newton Broyden-Fletcher-Goldfarb-Shanno method was initialized by a vector $\boldsymbol{v}^*$ of zeros and $\lambda = 0$ and run to the minimum length of gradient G. An empiric scaling factor of 10 of the default value of $\alpha = M$ can be used for putting more emphasis on either the accuracy of the mapping reconstruction (larger α) or on the minimum norm of $\boldsymbol{\omega}^*$ (smaller α).

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

2.2 Standardization of mapping coefficient vectors

Distance metric learning refers to finding a matrix $\mathbf{\Omega}$ such that a given criterion, like class discrimination or regression, is optimized for the mapped data relationships defined by

$$\mathrm{d}^{ij}_{p,\mathbf{\Omega}} = \mathrm{d}_{p,\mathbf{\Omega}}(\boldsymbol{x}^i, \boldsymbol{x}^j) = \left((\boldsymbol{x}^i - \boldsymbol{x}^j) \cdot \mathbf{\Omega} \cdot (\boldsymbol{x}^i - \boldsymbol{x}^j)^\mathsf{T}\right)^p, \quad p > 0. \tag{5}$$

Metrics are obtained for positive semi-definite matrices $\mathbf{\Omega}$, which is always the case for $\mathbf{\Omega} = \boldsymbol{\omega} \cdot \boldsymbol{\omega}^\mathsf{T}$, allowing unconstrained optimization of matrix $\boldsymbol{\omega}$. The two most common choices in Equation 5 are $p = 1$ for quadratic forms and $p = \frac{1}{2}$ for natural 'Mahalanobis' types of distances.

Since external mapping targets, such as the clustering of data in the mapped subspace, are invariant under rotations of $\mathbf{\Omega}$, a projection of $\boldsymbol{\omega}$ to the eigenvectors of $\mathbf{\Omega}$ helps to standardize the appearance of the columns of $\boldsymbol{\omega}$ [8]. Standardized directions are achieved by mirroring heavier tails of the eigenvector-rotated coefficient distributions into the positive half-space. This standardization of $\boldsymbol{\omega}$ does not change $\mathbf{\Omega}$, hence, all pairwise distances of mapped data remain unaffected.

2.3 Assessment of data attribute contributions to linear mappings

Although it is tempting to rate large absolute values of linear mapping coefficients as 'important', their proper interpretation is complicated by several aspects. For example, if correlated attributes exist in a data set, positively weighted ones can be compensated by their negative-weighted correlates. Also, in class separation tasks, two attributes might possess equal discriminative power, but at different variance levels, yet, the magnitude of coefficient weightings will be inversely related to variance. Sensitivity-based assessment is thus proposed.

Given vectors $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$, contributions of $\mathbf{\Omega}$ and $\boldsymbol{\omega}$ to the metric (5) are:

$$\frac{\partial \mathrm{d}^{ij}_{p,\mathbf{\Omega}}}{\partial \mathbf{\Omega}} = p \cdot (\boldsymbol{x}^i - \boldsymbol{x}^j)^\mathsf{T} \cdot (\boldsymbol{x}^i - \boldsymbol{x}^j) \cdot (\mathrm{d}^{ij}_{p,\mathbf{\Omega}})^{(p-1)/p}, \tag{6}$$

$$\frac{\partial \mathrm{d}^{ij}_{p,\mathbf{\Omega}}}{\partial \boldsymbol{\omega}} = 2 \cdot p \cdot \left((\boldsymbol{x}^i - \boldsymbol{x}^j) \cdot \boldsymbol{\omega}\right)^\mathsf{T} \cdot (\boldsymbol{x}^i - \boldsymbol{x}^j) \cdot (\mathrm{d}^{ij}_{p,\mathbf{\Omega}})^{(p-1)/p}. \tag{7}$$

Data vector relationships are encoded in their distance matrix, thus, the overall parameter contribution matrices depend on all involved data pairs:

$$\boldsymbol{V}_{p,\mathbf{\Omega}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial \mathrm{d}^{ij}_{p,\mathbf{\Omega}}}{\partial \mathbf{\Omega}} \quad \text{and} \quad \boldsymbol{V}_{p,\boldsymbol{\omega}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial \mathrm{d}^{ij}_{p,\mathbf{\Omega}}}{\partial \boldsymbol{\omega}}. \tag{8}$$

Note that usual quadratic forms related to $p = 1$ yield $\boldsymbol{V}_{1,\mathbf{\Omega}} = (N^2 - N) \cdot \mathsf{cov}(\mathbf{X})$, that is, the total derivative $\boldsymbol{V}_{1,\mathbf{\Omega}}$ is only depending on the covariance of the data, not on the mapping defined by $\mathbf{\Omega}$. Involvement of parameters is captured in $\boldsymbol{V}_{p,\boldsymbol{\omega}}$, particularly, $\boldsymbol{V}_{\frac{1}{2},\boldsymbol{\omega}}$ allows to assess the contribution of linear mapping coefficients $\boldsymbol{\omega}$ to relationships in Euclidean subspaces.

External models for computing the matrix $\boldsymbol{\omega}$ are likely to induce mostly non-vanishing derivatives resulting from Equation 7, and it might seem attractive to

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

use the column sums of squares of $\boldsymbol{V}_{p,\boldsymbol{\omega}}$ as an indicator of attribute importance. This measure is not very specific to the linear mapping though. In fact, all coefficient matrices that provide identical distance matrices of the projections yield identical $\boldsymbol{V}_{p,\boldsymbol{\omega}}$, because different $\boldsymbol{\Omega_1} = \boldsymbol{\omega_1} \cdot \boldsymbol{\omega_1}^\top$ and $\boldsymbol{\Omega_2} = \boldsymbol{\omega_2} \cdot \boldsymbol{\omega_2}^\top$ lead to

$$(\boldsymbol{x}^i - \boldsymbol{x}^j) \cdot \boldsymbol{\omega_1} = (\boldsymbol{x}^i - \boldsymbol{x}^j) \cdot \boldsymbol{\omega_2} \text{ [Eqn. 5]} \quad \rightarrow \quad \frac{\partial \mathrm{d}^{ij}_{p,\boldsymbol{\Omega_1}}}{\partial \boldsymbol{\omega_1}} = \frac{\partial \mathrm{d}^{ij}_{p,\boldsymbol{\Omega_2}}}{\partial \boldsymbol{\omega_2}} \text{ [Eqn. 7]}. \quad (9)$$

Therefore, a mapping-specific attribute assessment strategy is proposed.

By setting entries in row l of the mapping matrix to zero, the l-th data attributes are projected to zero and, therefore, ignored in calculations of distances in the mapped data. The gradient $\boldsymbol{V}^{l-}_{p,\boldsymbol{\omega}}$ related to each such hold-out attribute is used for calculating its difference to the original gradient $\boldsymbol{V}_{p,\boldsymbol{\omega}}$. Finally, the Frobenius norm of this gradient difference matrix, again ignoring formally abandoned contributions of row l, is assigned as sensitivity to the l-th attribute:

$$\mathsf{s}_l = \| \, [\boldsymbol{V}_{p,\boldsymbol{\omega}} - \boldsymbol{V}^{l-}_{p,\boldsymbol{\omega}}]^{l-} \, \|_F. \quad (10)$$

This differential view identifies parameters that yield distance distortion due to attribute hold-out compared to the original mapping.

The proposed strategy does not require a retraining of external models. Besides adding computing time, the alternative of retraining with a discarded attribute might lead to a phase transition in their cost function, and hence, to a completely different configuration with drastically changed contributions of the remaining attributes. As another benefit, the proposed way is expected to provide consistent results without even requiring access to external models.

3 Results

Wine sample spectra mappings are analyzed related to mid-infrared range scans at 256 contiguous wave numbers [11]. The mapping goal was the prediction of given alcohol concentrations. According to the work of [6] the spectra 34, 35 and 84 were considered as outliers and therefore discarded. The training and test sets contain 94 and 30 spectra, respectively, creating thus an under-determined regression problem which is considered as being solved by external models. Notice that the goodness of those models is subordinate to their interpretation here.

First, a least squares regression mapping is taken from Moore-Penrose pseudoinverse of the data matrix. The Pearson correlation between the predicted labels and the true labels is at maximum of 1 for the training data and of 0.990 for the 30 test samples. After application of the proposed regularization, the training set correlation decreased slightly to 0.999 while increasing to 0.995 for the test data. Corresponding mapping coefficients are shown in the top row of Figure 1, indicating a drastic difference between the very fluctuating raw coefficients (left) and the regularized ones with coherent spectral ranges (center).

Second, for further illustration coefficient vectors from correlative matrix mapping (CMM) to a 2-dimensional subspace are regularized and standardized,

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

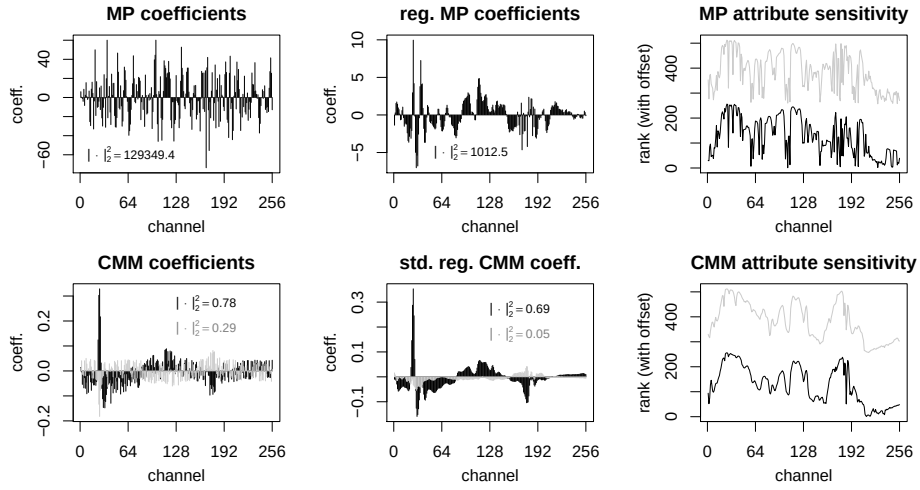

Fig. 1: Regularization and attribute assessment for 256-dimensional spectral data of wine samples linearly associated with alcohol content. **Left:** Raw mapping coefficient vectors from external methods with numbers indicating squared ℓ_2 norms. Top, Moore-Penrose pseudoinverse (MP). Bottom, 2-dimensional CMM regression subspace with mapping coefficient vectors in black and gray. **Center:** regularized coefficients, reflecting coherence of spectral channels of original dataset (not shown). Top, MP results. Bottom, CMM with regularized coefficients rotated on their eigenvectors. **Right:** Attribute relevance assessment based on ranks of differential sensitivity analysis (black lines) and on inverse ranks of mapping quality being integrated over all neighbor sizes [7] (gray lines, offset by 256). Higher ranks are more relevant.

shown in the bottom row of Figure 1. Two-banded fluctuations and similar amplitudes of both vectors in the left get coherent and separated into a high and low amplitude vector in the center plot. By definition of the involved operations, pairwise distances of mapped points are not affected despite strong changes of the mapping coefficient vectors. Regularized views even allow to detect similarities between CMM and Moore-Penrose solutions in the center column of Figure 1. According to the regularization target, values of coefficient vector norm drop from the left to the center column.

Third, attribute sensitivities are provided as ranks (higher is more important) in the right column of Figure 1 for the regularized Moore-Penrose and CMM mappings. A certain agreement between the two corresponding black lines can be observed at different degrees of fluctuation. For reference, results from a completely different method are included as gray lines, involving neighborhood ranks of projections for assessing the quality of mappings [7]. Excellent agreement at Spearman rank correlation values above 0.95 indicate very consistent attribute rankings between the methods for both cases, but the proposed way is computationally less demanding by a rough factor of $\mathcal{O}(n \cdot \log n)$, because method-inherent derivatives can be used instead of overall neighborhood ranking based on sorting operations.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

4 Conclusions

As exemplarily shown for a spectral data base, posterior norm-based regularization of under-determined linear mappings enables recovery of contiguous mapping coefficients representing smooth frequency ranges by using only the data and the original coefficients. The proposed regularization method resembles compressive sensing, but for computational tractability it currently makes use of ℓ_2 rather than ℓ_1 norm minimization. Original distance relationships in the mapping space are preserved after regularization. Rotation standardization helps to resolve further ambiguities. Finally, attribute hold-out for distance derivatives highlights data attributes which are most influential to the mapping. Software demonstrating improved generalization for additional examples from another food spectral database and from a colon cancer diagnosis task based on gene expression data is available as MATLAB/GNU-Octave package 'RegLin' at https://mloss.org/.

We thank Axel Soto, Dalhousie University, for his helpful remarks. This work is supported by the LOEWE Center for Synthetic Synthetic Microbiology (SYNMIKRO) and by grant XP3624HP/0606T from Saxony-Anhalt.

References

[1] *Multivariate Statistics*. General Books LLC, 2010.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul. On kernelization of supervised Mahalanobis distance learners. *Computing Research Repository (CoRR)*, pages 1–16, 2008.

[4] R. L. Gorsuch. *Factor Analysis*. Psychology Press, 2nd edition, 1983.

[5] R. Jin, S. Wang, and Y. Zhou. Regularized Distance Metric Learning: Theory and Algorithm. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 862–870. 2009.

[6] 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)*, pages 157–162. D-side Publications, 2007.

[7] J. A. Lee and M. Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31(14):2248–2257, 2010.

[8] P. Schneider, M. Biehl, and B. Hammer. Distance learning in discriminative vector quantization. *Neural Computation*, 21(10):2942–2969, 2009.

[9] M. Strickert, A. J. Soto, and G. E. Vazquez. Adaptive matrix distances aiming at optimum regression subspaces. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks (ESANN)*, pages 93–98. D-facto Publications, 2010.

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

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

[12] J. Ye, Z. Zhao, and H. Liu. Adaptive distance metric learning for clustering. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–7, 2007.