Publication III

III

Timo Similä and Jarkko Tikka (2005). Multiresponse sparse regression with application to multidimensional scaling, *in* W. Duch, J. Kacprzyk, E. Oja and S. Zadrozny (eds), *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005, 15th International Conference, Proceedings, Part II*, Vol. 3697 of *Lecture Notes in Computer Science*, Springer, pp. 97–102.

Multiresponse Sparse Regression with Application to Multidimensional Scaling

Timo Similä and Jarkko Tikka

Helsinki University of Technology, Laboratory of Computer and Information Science,
P.O. Box 5400, FI-02015 HUT, Finland
timo.simila@hut.fi
tikka@mail.cis.hut.fi

Abstract. Sparse regression is the problem of selecting a parsimonious subset of all available regressors for an efficient prediction of a target variable. We consider a general setting in which both the target and regressors may be multivariate. The regressors are selected by a forward selection procedure that extends the Least Angle Regression algorithm. Instead of the common practice of estimating each target variable individually, our proposed method chooses sequentially those regressors that allow, on average, the best predictions of all the target variables. We illustrate the procedure by an experiment with artificial data. The method is also applied to the task of selecting relevant pixels from images in multidimensional scaling of handwritten digits.

1 Introduction

Many practical data analysis tasks, for instance in chemistry [1], involve a need to predict several target variables using a set of regressors. Various approaches have been proposed to regression with a multivariate target. The target variables are often predicted separately using techniques like Ordinary Least Squares (OLS) or Ridge Regression [2]. An extension to multivariate prediction is the Curds and Whey procedure [3], which aims to take advantage of the correlational structure among the target variables. Latent variable models form another class with the same goal including methods like Reduced Rank, Canonical Correlation, Principal Components and Partial Least Squares Regression [4].

Prediction accuracy for novel observations depends on the complexity of the model. We consider only linear models, where the prediction accuracy is traditionally controlled by shrinking the regression coefficients toward zero [5,6]. In the latent variable approach the data are projected onto a smaller subspace in which the model is fitted. This helps with the curse of dimensionality but the prediction still depends on all of the regressors. On the contrary, sparse regression aims to select a relevant subset of all available regressors. Many automatic methods exist for the subset search including forward selection, backward elimination and various combinations of them. Least Angle Regression (LARS) [7] is a recently introduced algorithm that combines forward selection and shrinkage.

The current research in sparse regression is mainly focused on estimating a univariate target. We propose a Multiresponse Sparse Regression (MRSR) algo-

W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 97–102, 2005.

rithm, which is an extension of the LARS algorithm. Our method adds those regressors sequentially to the model, which allow the most accurate predictions averaged over all the target variables. This allows to assess the average importance of the regressors in the multitarget setting. We illustrate the MRSR algorithm by artificially generated data and also apply it in a discriminative projection of images representing handwritten digits.

2 Multiresponse Sparse Regression

Suppose that the targets are denoted by an $n \times p$ matrix $\boldsymbol{T} = [\boldsymbol{t}_1 \cdots \boldsymbol{t}_p]$ and the regressors are denoted by an $n \times m$ matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_m]$. The MRSR algorithm adds sequentially active regressors to the model

$$\boldsymbol{Y}^k = \boldsymbol{X}\,\boldsymbol{W}^k \tag{1}$$

such that the $n \times p$ matrix $\boldsymbol{Y}^k = [\boldsymbol{y}_1^k \cdots \boldsymbol{y}_p^k]$ models the targets $\boldsymbol{T}$ appropriately. The $m \times p$ weight matrix $\boldsymbol{W}^k$ includes k nonzero rows in the beginning of the kth step. Each step introduces a new nonzero row and, thus, a new regressor to the model. In the case $p = 1$ MRSR coincides with the LARS algorithm. This makes MRSR rather an extension than an improvement of LARS.

Set $k = 0$, initialize all elements of $\boldsymbol{Y}^0$ and $\boldsymbol{W}^0$ to zero, and normalize both $\boldsymbol{T}$ and $\boldsymbol{X}$ to zero mean. The columns of $\boldsymbol{T}$ and the columns of $\boldsymbol{X}$ should also have the same scales, which may differ between the matrices. Define a cumulative correlation between the jth regressor $\boldsymbol{x}_j$ and the current residuals

$$c_j^k = ||(\boldsymbol{T} - \boldsymbol{Y}^k)^T \boldsymbol{x}_j||_1 = \sum_{i=1}^{p} |(\boldsymbol{t}_i - \boldsymbol{y}_i^k)^T \boldsymbol{x}_j|. \tag{2}$$

The criterion measures the sum of absolute correlations between the residuals and the regressor over all p target variables in the beginning of the kth step. Let the maximum cumulative correlation be denoted by $c_{\max}^k$ and the group of regressors that satisfy the maximum by $\mathcal{A}$, or formally

$$c_{\max}^k = \max_j \{c_j^k\}, \quad \mathcal{A} = \{j \mid c_j^k = c_{\max}^k\}. \tag{3}$$

Collect the regressors that belong to $\mathcal{A}$ as an $n \times |\mathcal{A}|$ matrix $\boldsymbol{X}_{\mathcal{A}} = [\cdots \boldsymbol{x}_j \cdots]_{j \in \mathcal{A}}$ and calculate an OLS estimate

$$\bar{\boldsymbol{Y}}^{k+1} = \boldsymbol{X}_{\mathcal{A}} (\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}})^{-1} \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{T}. \tag{4}$$

Note that the OLS estimate involves $k + 1$ regressors at the kth step.

Greedy forward selection adds regressors based on (2) and the OLS estimate (4) is used. However, we define a less greedy algorithm by moving from the MRSR estimate $\boldsymbol{Y}^k$ toward the OLS estimate $\bar{\boldsymbol{Y}}^{k+1}$, i.e. in the direction $\boldsymbol{U}^k = \bar{\boldsymbol{Y}}^{k+1} - \boldsymbol{Y}^k$, but we will not reach it. The largest step possible is taken

in the direction of $\boldsymbol{U}^k$ until some $\boldsymbol{x}_j$, where $j \notin \mathcal{A}$, has as large cumulative correlation with the current residuals as the already added regressors [7]. The MRSR estimate $\boldsymbol{Y}^k$ is updated

$$\boldsymbol{Y}^{k+1} = \boldsymbol{Y}^k + \gamma^k(\bar{\boldsymbol{Y}}^{k+1} - \boldsymbol{Y}^k). \tag{5}$$

In order to make the update, we need to calculate the correct step size γ^k. The cumulative correlations c_j^{k+1} may be obtained by substituting (5) into (2). According to (4), we may write $\boldsymbol{X}_{\mathcal{A}}^T(\bar{\boldsymbol{Y}}^{k+1} - \boldsymbol{Y}^k) = \boldsymbol{X}_{\mathcal{A}}^T(\boldsymbol{T} - \boldsymbol{Y}^k)$. This gives the cumulative correlations in the next step as a function of γ

$$c_j^{k+1}(\gamma) = |1 - \gamma| c_{\max}^k \text{ for all } j \in \mathcal{A} \tag{6}$$

$$c_j^{k+1}(\gamma) = ||\boldsymbol{a}_j^k - \gamma \boldsymbol{b}_j^k||_1 \text{ for all } j \notin \mathcal{A}, \tag{7}$$

where $\boldsymbol{a}_j^k = (\boldsymbol{T} - \boldsymbol{Y}^k)^T \boldsymbol{x}_j$ and $\boldsymbol{b}_j^k = (\bar{\boldsymbol{Y}}^{k+1} - \boldsymbol{Y}^k)^T \boldsymbol{x}_j$. A new regressor with index $j \notin \mathcal{A}$ will enter the model when (6) and (7) are equal. This happens if the step size is taken from the set

$$\Gamma_j = \left\{ \frac{c_{\max}^k + \boldsymbol{s}^T \boldsymbol{a}_j^k}{c_{\max}^k + \boldsymbol{s}^T \boldsymbol{b}_j^k} \right\}_{\boldsymbol{s} \in \mathcal{S},} \tag{8}$$

where $\mathcal{S}$ is the set of all 2^p sign vectors of size $p \times 1$, i.e. the elements of $\boldsymbol{s}$ may be either 1 or -1. The correct choice is the smallest of such positive step sizes that introduces a new regressor

$$\gamma^k = \min\{\gamma \mid \gamma \geq 0 \text{ and } \gamma \in \Gamma_j \text{ for some } j \notin \mathcal{A}\}, \tag{9}$$

which completes the update rule (5).

The weight matrix, which satisfies (5) and (1) may be updated

$$\boldsymbol{W}^{k+1} = (1 - \gamma^k)\, \boldsymbol{W}^k + \gamma^k\, \bar{\boldsymbol{W}}^{k+1}, \tag{10}$$

where $\bar{\boldsymbol{W}}^{k+1}$ is an $m \times p$ sparse matrix. Its nonzero rows, which are indexed by $j \in \mathcal{A}$, contain the corresponding rows of the OLS parameters $(\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}})^{-1} \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{T}$. The parameters of the selected regressors are shrunk according to (10) and the rest are kept at zero during the steps $k = 0, \ldots, m-2$. The last step coincides with the OLS parameters. The selection of the final model from m possibilities is based on prediction accuracy for novel data.

3 Multidimensional Scaling

Multidimensional scaling (MDS) [8] is a collection of techniques for exploratory data analysis that visualize proximity relations of objects as points in a low-dimensional Euclidean feature space. Proximities are represented as pairwise dissimilarity values δ_{ij}. We concentrate on the Sammon criterion [9]

$$E(\boldsymbol{Y}) = \sum_{i=1}^{n} \sum_{j>i} \alpha_{ij} (||\hat{\boldsymbol{y}}_i - \hat{\boldsymbol{y}}_j||_2 - \delta_{ij})^2. \tag{11}$$

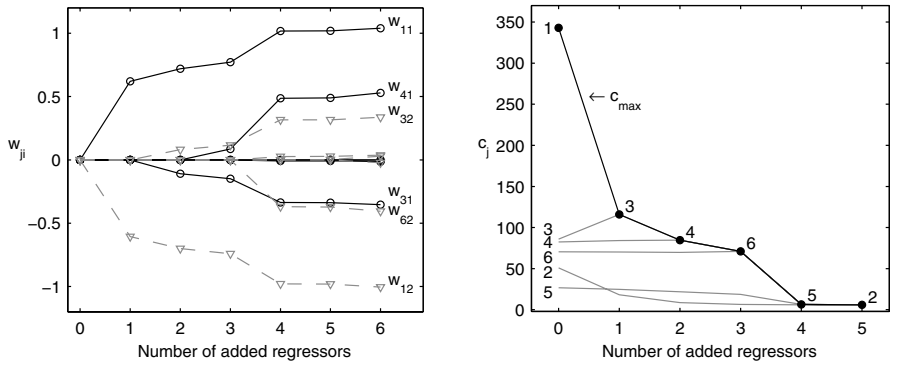

Fig. 1. Results for the artificial data: (*left*) Estimates of the weights w_{ji} and (*right*) cumulative correlations c_j as a function of the number of regressors in the model

Normalization coefficients $\alpha_{ij} = 2/(n(n-1)\delta_{ij})$ put focus on similar objects. The vector $\hat{\boldsymbol{y}}_i$ is the ith row of an $n \times p$ feature configuration matrix $\boldsymbol{Y}$.

Differing from the ordinary Sammon mapping, we are not only seeking for $\boldsymbol{Y}$ that minimizes the error (11), but also the parameterized transformation from the data space to the feature space that generates $\boldsymbol{Y}$ as a function of an $n \times m$ matrix $\boldsymbol{X}$. More specifically, we define $\boldsymbol{Y}$ as a linear combination of some relevant columns of $\boldsymbol{X}$. Next, a gradient descent procedure for such a minimization of (11) is outlined by modifying the Shadow Targets algorithm [10].

Make an initial guess $\boldsymbol{Y}^0$ and set the learning rate parameter η^0 to a small positive value. The estimated targets at each of the following iterations are

$$\boldsymbol{T}^{\ell+1} = \boldsymbol{Y}^{\ell} - \eta^{\ell} \frac{\partial E(\boldsymbol{Y}^{\ell})}{\partial \boldsymbol{Y}}. \tag{12}$$

Calculate $\boldsymbol{W}^{\ell+1}$ by feeding $\boldsymbol{T}^{\ell+1}$ and $\boldsymbol{X}$ to the MRSR algorithm and update $\boldsymbol{Y}^{\ell+1} = \boldsymbol{X}\boldsymbol{W}^{\ell+1}$. As suggested in [10], set $\eta^{\ell+1} = 0.1\eta^{\ell}$ if error (11) has increased from the previous iteration, and otherwise set $\eta^{\ell+1} = 1.2\eta^{\ell}$.

The only difference between the original Shadow Targets algorithm is the way in which the weights $\boldsymbol{W}^{\ell+1}$ are calculated. MRSR replaces the calculation of OLS parameters $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{T}^{\ell+1}$. This allows us to control the sparsity of the solution. The number of nonzero rows in $\boldsymbol{W}^{\ell+1}$ depends on the number of steps we perform in the MRSR algorithm.

4 Experiments

To illustrate the MRSR algorithm, we generated artificial data from the setting $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{E}$, where the elements of a 200×6 matrix $\boldsymbol{X}$ are independently distributed according to the Gaussian distribution $N(0,1)$, the elements of a 200×2 matrix $\boldsymbol{E}$ according to $N(0, 0.35^2)$, and the weights are set to

$$\boldsymbol{W}^T = \begin{bmatrix} 1 & 0 & -1/3 & 1/2 & 0 & 0 \\ -1 & 0 & 1/3 & 0 & 0 & -2/5 \end{bmatrix}.$$

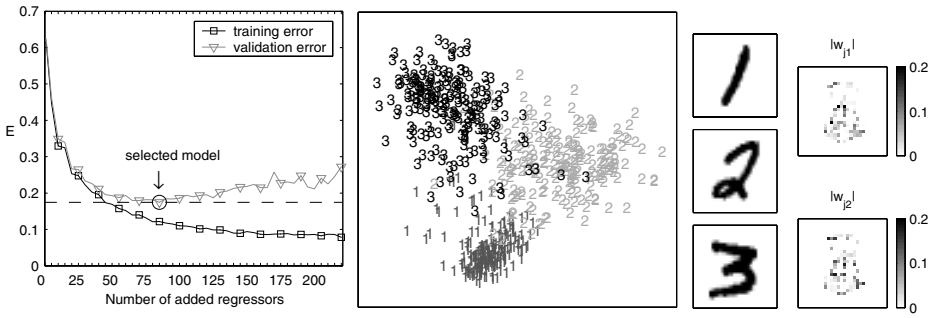

Fig. 2. Results for the digits data: (*left*) Training and validation errors as a function of the number of regressors in the model. (*middle*) Projection of the test set. (*right*) Example images from the test set and images illustrating the weights w_{ji}.

Fig. 1 shows results of MRSR analysis of the artificial data. The regressors are added to the model in the order 1, 3, 4, 6, 5, 2 and each addition decreases the maximum cumulative correlation between the regressors and residuals. The apparently most important regressor $\boldsymbol{x}_1$ is added first and the two useless regressors $\boldsymbol{x}_2$ and $\boldsymbol{x}_5$ last. Importantly, $\boldsymbol{x}_3$ enters the model before $\boldsymbol{x}_4$ and $\boldsymbol{x}_6$, because it is overall more relevant. However, $\boldsymbol{x}_4$ ($\boldsymbol{x}_6$) would enter before $\boldsymbol{x}_3$ if the first (second) target was estimated individually using the LARS algorithm.

The second experiment studies images of handwritten digits 1, 2, 3 with 28×28 resolution from the MNIST database. An image is represented as a row of $\boldsymbol{X}$, which consists of grayscale values of 784 pixels between zero and one. We constructed randomly three distinct data sets: a training set with 100, validation set with 200, and test set with 200 samples per digit. The aim is to form a model that produces a discriminative projection of the images onto two dimensions by a linear combination of relevant pixels. Pairwise dissimilarities are calculated using a discriminative kernel [11]. A within digit dissimilarity is $\delta_{ij} = 1 - \exp(-||\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j||^2/\beta)$ and a between digit dissimilarity is $\delta_{ij} = 2 - \exp(-||\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{x}}_j||^2/\beta)$, where $\hat{\boldsymbol{x}}_i$ denotes the ith image. The parameter β controls discrimination and we found a visually suitable value $\beta = 150$.

Fig. 2 displays results for the digits data. The left panel shows the best training set error of MDS starting from 100 random initializations $\boldsymbol{Y}^0$ and the corresponding validation error as a function of the number of effective pixels in the model. The validation error is at the minimum when the model uses 85 pixels. The middle panel shows a projection of test images obtained by this model and the right panel illustrates sparsity of the estimated weights w_{ji}. The selected group of about 11% of the pixels is apparently enough to form a successful linear projection of novel images.

5 Conclusions

We have presented the MRSR algorithm for forward selection of regressors in the estimation of a multivariate target using a linearly parameterized model. The

algorithm is based on the LARS algorithm, which is designed for a univariate target. MRSR adds regressors one by one to the model such that the added regressor always correlates most of all with the current residuals. The order in which the regressors enter the model reflects their importance. Sparsity places focus on relevant regressors and makes the results more interpretable. Moreover, sparsity coupled with shrinkage helps to avoid overfitting.

We used the proposed algorithm in an illustrative experiment with artificially generated data. In another experiment we studied images of handwritten digits. The algorithm fitted a MDS model that allows a discriminative projection of the images onto two dimensions. The experiment combines the two major categories of dimensionality reduction methods: input selection and input transformation.

LARS is closely connected to the Lasso estimator [6,7]. As such, MRSR does not implement a multiresponse Lasso, which constraints the ℓ_1-norm of the weight matrix. MRSR updates whole rows of the matrix instead of its individual elements. However, the connection might emerge by modifying the constraint structure of Lasso. Another subject of future research could be basis function selection for linear neural networks.

References

1. Burnham, A.J., MacGregor, J.F., Viveros, R.: Latent Variable Multivariate Regression Modeling. Chemometrics and Intelligent Laboratory Systems **48** (1999) 167-180
2. Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics **12** (1970) 55-67
3. Breiman, L., Friedman, J.H.: Predicting Multivariate Responses in Multivariate Regression. Journal of the Royal Statistical Society. Series B **59** (1997) 3-54
4. Abraham, B., Merola, G.: Dimensionality Reduction Approach to Multivariate Prediction. Computational Statistics & Data Analysis **48** (2005) 5-16
5. Copas, J.: Regression, Prediction and Shrinkage. Journal of the Royal Statistical Society. Series B **45** (1983) 311-354
6. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B **58** (1996) 267-288
7. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least Angle Regression. Annals of Statistics **32** (2004) 407-499
8. Cox, T.F., Cox, M.A.A.: Multidimensional Scaling. Monographs on Statistics and Applied Probability 88. Chapman & Hall (2001)
9. Sammon, J.W.: A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers **C-18** (1969) 401-409
10. Tipping, M.E., Lowe, D.: Shadow Targets: A Novel Algorithm for Topographic Projections by Radial Basis Functions. Neurocomputing **19** (1998) 211-222
11. Zhang, Z.: Learning Metrics via Discriminant Kernels and Multidimensional Scaling: Toward Expected Euclidean Representation. In: International Conference on Machine Learning. (2003) 872-879