This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence



# Multi-Target Regression via Robust Low-Rank Learning

Xiantong Zhen, Mengyang Yu, *Student Member, IEEE*, Xiaofei He, *Senior Member, IEEE*, and Shuo Li

**Abstract**—Multi-target regression has recently regained great popularity due to its capability of simultaneously learning multiple relevant regression tasks and its wide applications in data mining, computer vision and medical image analysis, while great challenges arise from jointly handling inter-target correlations and input-output relationships. In this paper, we propose Multi-layer Multi-target Regression (MMR) which enables simultaneously modeling intrinsic inter-target correlations and nonlinear input-output relationships in a general framework via robust low-rank learning. Specifically, the MMR can explicitly encode inter-target correlations in a structure matrix by matrix elastic nets (MEN); the MMR can work in conjunction with the kernel trick to effectively disentangle highly complex nonlinear input-output relationships; the MMR can be efficiently solved by a new alternating optimization algorithm with guaranteed convergence. The MMR leverages the strength of kernel methods for nonlinear feature learning and the structural advantage of multi-layer learning architectures for inter-target correlation modeling. More importantly, it offers a new multi-layer learning paradigm for multi-target regression which is endowed with high generality, flexibility and expressive ability. Extensive experimental evaluation on 18 diverse real-world datasets demonstrates that our MMR can achieve consistently high performance and outperforms representative state-of-the-art algorithms, which shows its great effectiveness and generality for multivariate prediction.

**Index Terms**—Robust Low-Rank Learning, Multi-Layer Learning, Multi-Target Regression, Matrix Elastic Nets.

## 1 Introduction

Multi-target regression, as an instance of multitask learning [1], has recently drawn increasing research efforts in the machine learning community due to its great capability of predicting multiple relevant targets simultaneously with improved performance. Moreover, it has started to show its great effectiveness to solve challenging problems in a broad range of applications including data mining [2], computer vision [3] and medical image analysis [4].

The major challenges of multi-target regression arise from jointly modeling inter-target correlations and nonlinear input-output relationships [5]. By exploring the shared knowledge across relevant targets to capture the inter-target correlation, multi-target regression performance can be significantly improved [6], [7]. However, the structure of inter-target correlations is usually not known *a priori* and varies greatly with different applications. Meanwhile, multiple targets represent higher-level semantic concepts of high-dimensional inputs [8], [9], which induces highly nonlinear relationships between inputs and targets. Although great effort has been made to improve multi-target regression in the last decades [10], it still lacks a general framework that can well tackle those two major challenges simultaneously.

To explore inter-target correlations, most of existing multi-target regression models were focused on designing a regularizer on the regression matrix, which rely either on linear regression models [11], [12], [13] or on specific assumptions of correlation structures with strong prior knowledge [14], [15], [16]. By building on linear regression, sparsity or low rank was simply imposed on the regression matrix to explore the correlations. However, these linear models lack ability to handle nonlinear input-output relationships; moreover it is nontrivial to extend for nonlinear regression due to the non-convexity of loss functions or the non-smoothness of sparsity constraints [17], [18]. Under specific assumptions, e.g., task parameters share a common prior [19], [20], or combine a finite number of basis tasks [21], particular structures of inter-target correlations were explored. However, those assumptions would not necessarily hold or be shared across different applications due to the great diversity of inter-target correlations in different domains, which resulting in models of insufficient generality to automatically extract the correlations from data for diverse applications [5].

To handle nonlinear input-output relationships, kernel methods [22], [23] were extended from single task learning to multi-task learning, which however do not provide effective ways to model the inter-

- *X. Zhen and S. Li are with the Department of Medical Biophysics, The University of Western Ontario, London, Ontario, N6A 4V2, Canada. E-mail: zhenxt@gmail.com, slishuo@gmail.com*
- *M. Yu is with the Department of Computer and Information Science, Northumbria University, Newcastle upon Tyne, NE1 8ST, U.K. E-mail: m.y.yu@ieee.org*
- *X. He is with the State Key Lab of CAD&CG, Zhejiang University, Hangzhou, Zhejiang, China, 310058. E-mail: xiaofeihe@cad.zju.edu.cn*

0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence



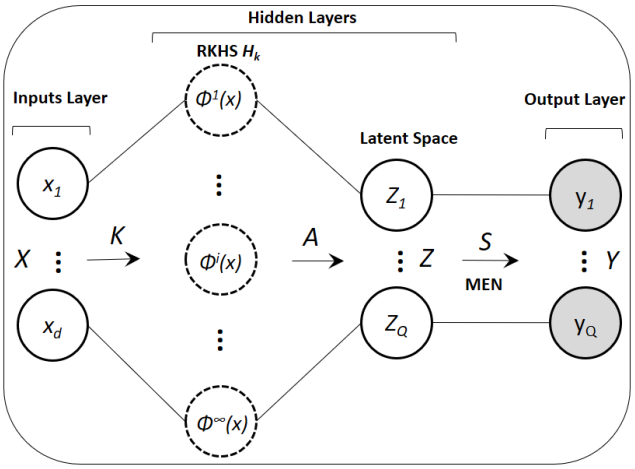


Fig. 1: The architecture of the proposed multi-layer multi-target regression (MMR).

target correlation. In [23], for instance, the regression matrix of multiple tasks is simply reshaped into a vector to explore inter-target correlations, which does not distinguish between inter and intra tasks, and tends to be less effective to encode the correlations. Although kernel extension was developed in multi-target relationship learning (MTRL) [5], a matrix-variate normal distribution is required as a prior to model task structure in a covariance matrix [13].

In this paper, we propose Multi-layer Multi-target Regression (MMR) that enables simultaneously modeling intrinsic inter-target correlations and complex input-output relationships in a general framework. As illustrated in Fig. 1, the MMR accomplishes a multi-layer learning architecture composed of the input, hidden and output (target) layers.

- The high-dimensional inputs $X$ are implicitly mapped into a high, even infinite dimensional reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\mathcal{K}$ induced by some nonlinear kernel $\mathcal{K}$; the mapping serves as a nonlinear feature extraction function, which allows to disentangle highly nonlinear input-output relationships.
- The variables $Z$ in the latent space, which are obtained by a linear transformation $A$ via a representer theorem, represents higher-level concepts to build a common feature representation for multiple targets [8]; the latent space decouples inputs and targets, which allows to effectively handle their different noise levels by $A$ and $S$, respectively [24].
- The structure matrix $S$ explicitly encodes the inter-target correlation by imposing the matrix elastic net (MEN) regularization [25], which enables robust low-rank learning of the correlation; $S$ is learned automatically from data without relying on any specific assumptions, which greatly enhances its generality.

The proposed MMR leverages the strength of kernel methods for nonlinear feature learning and the structural advantage of multi-layer architectures to capture inter-target correlations [26], and more importantly, it provides a new multi-layer learning paradigm that is endowed with high generality, flexibility and expressive ability for multi-target regression. Moreover, the MMR is highly extensible and can work with deep learning architectures, e.g., convolutional neural networks (CNNs) [27], for further fine-tuning [28].

**Contributions.** The contributions of this work are summarized as follows:

- We propose a new multi-layer multi-target regression model, which enables simultaneously modeling intrinsic inter-target correlations and complex input-output relationships in one single framework.
- We introduce the matrix elastic nets (MEN) to multi-target regression to explore inter-target correlations, which enables explicitly encoding the correlations by a robust low-rank learning framework without specific assumptions.
- We provide a kernel extension of the learning framework, which enables effectively disentangling highly nonlinear relationships between high-dimensional inputs and multiple targets.
- We derive a new alternating optimization algorithm, which enables efficiently solving the objective with quick convergence to achieve efficient multi-target regression.

## 2 Related Work

Multi-target regression [10] has recently regained great popularity due to its fundamental role in machine learning and widespread applications in computer vision and medical image analysis. Previous work has been focused on particular aspects, e.g., simply learning features for multiple tasks [7], [14], [29], [30] or solely exploring specific task structures [6], [31], [32], [33]. Some of the regression models are designed for classification tasks [12], [34]. Most of existing methods are developed mainly on linear regression models to explore inter-target correlations, while being lack of the ability to simultaneously handle nonlinear relationships between high-dimensional inputs and multiple targets.

Recently, Rai et al. [13] proposed multi-output regression with output and task structures (MROTS) which generalizes the multivariate regression model with covariance estimation (MRCE) [11] and the linear multi-task relationship learning (MTRL) [5] as its special cases with improved performance. However, like the MRCE [11], the MROTS does not provide any formulation for nonlinear regression. Liu et al. [35], [17] proposed a linear regression model called calibrated multivariate regression (CMR) to tackle different noise levels of different tasks. By assuming an uncorrelated

0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence



structure of noises, the CMR employs a $\ell_{2,1}$-norm based loss function to calibrate each regression task, while the assumption of uncorrelated noises does not always hold in practice. Beside, it is nontrivial to extend the CMR to kernel regression due to the $\ell_{2,1}$-norm loss function.

In order to handle nonlinear input-output relationship, the recent output kernel learning (OKL) algorithm [22], [36], which learns a semi-definite similarity matrix, i.e., the output kernel, of multiple targets, would not fully capture inter-target correlations, e.g., negative correlations [5]. By assuming that all tasks can be clustered into disjoint groups, the clustered multi-target learning (CMTL) [15] was developed to explore inter-target correlations, which learns the underlying cluster structure from the training data. However, the number of clusters needs to be specified, which is rarely available in real-world tasks. Recently, an improved version of CMTL called flexible clustered multi-target (FCMTL) was presented in [33]. In the FCMTL, the cluster structure is learned by identifying representative tasks. However, the assumption of the existence of representative tasks remains purely heuristic and therefore would not be shared across diverse applications.

## 3 Multi-Layer Multi-Target Regression

The proposed MMR accomplishes a general framework of multi-layer learning to jointly model inter-target correlations by robust low rank learning via matrix elastic nets (MEN) (Sec. 3.2) and disentangle nonlinear input-output relationships by kernel regression (Sec. 3.3); the MMR is efficiently solved by a newly derived alternating optimization algorithm with guaranteed convergence (Sec. 3.4).

### 3.1 Problem Formulation

We start with the fundamental multi-target linear regression model $\mathbf{y} = W\mathbf{x} + \mathbf{b}$, where $\mathbf{y} = [y_1, \ldots, y_i, \ldots, y_{\mathcal{Q}}]^\top \in \mathbb{R}^{\mathcal{Q}}$ are the multivariate targets, $\mathbf{x} \in \mathbb{R}^d$ is the input, $W = [\mathbf{w}_1, \ldots, \mathbf{w}_i, \ldots, \mathbf{w}_{\mathcal{Q}}]^\top \in \mathbb{R}^{\mathcal{Q}\times d}$ is the model parameter, i.e., the regression coefficient, each $\mathbf{w}_i \in \mathbb{R}^d$ is the predictor for $y_i$, $\mathbf{b} \in \mathbb{R}^{\mathcal{Q}}$ is the bias, $d$ and $\mathcal{Q}$ are the dimensionality of input and output spaces, respectively. Given the training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\mathcal{N}}$, one can solve for $W$ by solving the following penalized optimization objective:

$$W^* = \arg\min_W \frac{1}{\mathcal{N}}||Y - WX - B||_F^2 + \lambda||W||_F^2, \quad (1)$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathcal{N}}]$, $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{\mathcal{N}}]$, $B = [\ldots, \mathbf{b}, \ldots] \in \mathbb{R}^{Q\times N}$, $||A||_F^2$ is the Frobenius norm of matrix $A$, which can be computed by $tr(A^\top A)$, and $\lambda$ is the regularization parameter that controls the amount of shrinkage, i.e., the larger the value of $\lambda$, the greater the amount of shrinkage [37].

The objective function in (1) is a straightforward extension of ridge regression, one of classical statistical learning algorithms [37], to multivariate targets. When working in a reproducing kernel Hilbert space (RKHS), the resulting model is known as kernel ridge regression (KRR) [37] [38]. Likewise, (1) can be kernelized to achieve multi-target kernel ridge regression (mKRR). We derive the proposed MMR from this fundamental formulation to ensure its generality.

The multi-target regression model in (1) is decoupled into several single-output problems, which does not take into account inter-target correlations, resulting in suboptimal multi-target regression with inferior performance. In what follows, we introduce our MMR, which is a multi-layer learning architecture to explicitly model the correlations by a robust low-rank learning with matrix elastic nets (MEN).

### 3.2 Robust Low-Rank Learning with MEN

Rather than directly imposing sparsity on $W$ in existing methods, we propose incorporating a latent space, from which a structure matrix $S$ is learned to explicitly encode inter-target correlations via a rank minimization.

$$\min_{W,S} \frac{1}{\mathcal{N}}||Y - SZ||_F^2 + \lambda||W||_F^2 + \beta\mathrm{Rank}(S) + \gamma||S||_F^2, \quad (2)$$

where $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_i, \ldots, \mathbf{z}_{\mathcal{N}}] \in \mathbb{R}^{\mathcal{Q}\times\mathcal{N}}$, $\mathbf{z}_i = W\mathbf{x}_i + \mathbf{b} \in \mathbb{R}^{\mathcal{Q}}$ contains the latent variables in the latent space, $S \in \mathbb{R}^{\mathcal{Q}\times\mathcal{Q}}$ is the structure matrix that serves to explicitly model inter-target correlations, $\beta$ is the regularization parameter to control the rank of $S$, that is, a larger $\beta$ induces lower rank, and the Frobenius norm control the shrinkage of $S$ with the associated parameter $\gamma$. The rank minimization of the structure matrix $S$ explores the low-rank structure existing between tasks to capture the intrinsic inter-target correlation. $S$ is learned automatically from data without relying on any specific assumptions, which allows to adaptively cater different applications.

However, the objective function in (2) is NP-hard due to the noncontinuous and non-convex nature of the rank function [39]. The nuclear norm $||S||_*$ is commonly used and has been proven to be the convex envelop of the rank function over the domain $||S||_2 \leq 1$ [40], which provides the tightest lower bound among all convex lower bounds of the rank function $\mathrm{Rank}(S)$.

As a consequence, the combination of the nuclear norm with the Frobenius norm on $S$ gives rise to the matrix elastic net (MEN) [25] as a regularizer of (2):

$$\min_{W,S} \frac{1}{\mathcal{N}}||Y - SZ||_F^2 + \lambda||W||_F^2 + \beta||S||_* + \gamma||S||_F^2, \quad (3)$$

where the nuclear norm $||S||_*$ also known as the trace norm is a particular instance of the Schatten $p$-norm

0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence



[41] with $p = 1$, i.e., Schatten 1-norm. The Schatten p-norm of a matrix $M$ is defined as

$$||M||_{S_p} = \Big(\sum_i \sigma_i^p(M)\Big)^{1/p}, \tag{4}$$

where $0 < p \leq 2$, $\sigma_i(M)$ is the $i$-th largest singular value of $M$. When $p < 2$, the Schatten $p$-norm encourages sparsity of the singular values, which achieves rank minimization. With (4), the nuclear norm of $S$ can be written as $||S||_* = tr(\sqrt{S^\top S})$. It is interesting to mention that when $p = 0$, the Schatten 0-norm is defined as

$$||M||_{S_0} = \sum_i \sigma_i^0 \tag{5}$$

which is exactly the rank of $A$.

The MEN is an analog to the elastic-net regularization [42] from compressive sensing and sparse representation [42]. It has been shown that the elastic net often outperforms the lasso [42]. In the MEN, the nuclear-norm constraint enforces the low-rank property of the solution $S$ to encode inter-target correlations, and the Frobenius-norm constraint induces a linear shrinkage on the matrix entries leading to stable solutions [25]. The MEN regularization generalizes the matrix lasso and provides improved performance than lasso [43].

To the best of our knowledge, this is the first work that introduces the MEN to multi-target regression for robust low-rank learning, which offers a general framework to encode inter-target correlations. We highlight that the proposed multi-layer multi-target regression enjoys favorable merits:

- The latent space enables decouples inputs and targets from distinctive distributions, which allows them to be effectively handled respectively by the regression coefficient $W$ and the structure matrix $S$ [24], [44].
- The structure matrix which is detached from inputs by the latent space is able to effectively calibrate multiple targets according to their different noise levels to achieve optimal parameter estimation [35], [45].
- The matrix elastic nets (MEN) provide a general regularization network to achieve robust low-rank learning of the intrinsic inter-target correlations [25].

### 3.3 Kernel Extension

Due to the constraints directly imposed on the regression matrix $W$, most of the existing methods would not be kernelized for nonlinear regression. On the contrary, thanks to the multi-layer learning architecture, the MMR is more flexible and extensible, and naturally admits the Representer Theorem [46], [18] as shown in Theorem 1, which enables kernel extension to achieve nonlinear regression.

**Theorem 1.** *Given any fixed matrix $S$, the objective function in (3) w.r.t. $W$ which is defined over a Hilbert space $\mathcal{H}$. If (3) has a minimizer w.r.t. $W$, it admits a linear representer theorem of the form $W = AX^\top$, where $A \in \mathbb{R}^{\mathcal{Q}\times\mathcal{N}}$ is the coefficient matrix.*

*Remark.* Theorem 1 provides important theoretical guarantees for kernel extension to achieve nonlinear multi-target regression. The proof of the theorem is straightforward and omitted here due to space limit.

Based on Theorem 1, we can derive kernel extension in the reproducing kernel Hilbert space (RKHS) for kernel regression to handle nonlinear input-output relationships. To facilitate the derivation, we rewrite the objective function (3) in term of traces as follows:

$$\begin{aligned}\min_{W,S} \frac{1}{\mathcal{N}} tr((Y - S(WX + B))^\top (Y - S(WX + B))) \\ + \lambda tr(W^\top W) + \beta tr(\sqrt{S^\top S}) + \gamma tr(S^\top S).\end{aligned} \tag{6}$$

According to the linear representer theorem in Theorem 1, we have

$$W = A\Phi(X)^\top, \tag{7}$$

where $\Phi(X) = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_i), \ldots, \phi(\mathbf{x}_\mathcal{N})]$ and $A = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_i, \ldots, \boldsymbol{\alpha}_Q]^\top \in R^{\mathcal{Q}\times\mathcal{N}}$, $\boldsymbol{\alpha}_i \in \mathbb{R}^N$, and $\phi(\cdot)$ denotes the feature map of $\mathbf{x}_i$, which maps $\mathbf{x}_i$ to $\phi(\mathbf{x}_i)$ in some RKHS of high, even infinite dimensionality. The mapping serves as a nonlinear feature extraction to handle complicated input-target relationships. The corresponding kernel function $k(\cdot,\cdot)$ satisfies $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$.

Substituting (7) into (6), we obtain the following objective function:

$$\begin{aligned}\min_{A,S} \frac{1}{\mathcal{N}} tr((Y - SA\Phi(X)^\top \Phi(X))^\top (Y - SA\Phi(X)^\top \Phi(X))) \\ + \lambda tr((A\Phi(X)^\top)^\top (A\Phi(X)^\top)) + \beta tr(\sqrt{S^\top S}) \\ + \gamma tr(S^\top S).\end{aligned} \tag{8}$$

where the bias $B$ is omitted for simplicity since it has been proven that the bias can be absorbed into the regression coefficient $W$ by adding an additional dimension into input features $\mathbf{x}_i$ [47], [48].

We accomplish the kernel version of the multi-layer multi-target regression with the kernel matrix $K = \Phi(X)^\top \Phi(X)$ defined in the RKHS space.

$$\begin{aligned}\min_{A,S} \frac{1}{\mathcal{N}} tr((Y - SAK)^\top (Y - SAK)) + \lambda tr(AKA^\top) \\ + \beta tr(\sqrt{S^\top S}) + \gamma tr(S^\top S).\end{aligned} \tag{9}$$

In (9), the induced latent variables $Z = AK$ can extract high-level representations for multiple semantic targets, which allows to disentangle the nonlinear relationship between low-level inputs and semantic-level targets [8]. The latent space with high-level features will also facilitate the efficient linear low-rank learning of $S$ to model inter-target correlations [49]

0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence



to achieve more accurate multi-target prediction. The MMR in (9) leverages the strength of kernel methods for nonlinear feature extraction and the structural advantage of multi-layer architectures for inter-target correlation modeling. In contrast to existing multi-target regression models, the obtained MMR in (9) accomplishes a new multi-layer learning architecture, which is endowed with great generality, flexibility and expressive ability for diverse challenging tasks.

**Generality.** One of the important advantages of the proposed MMR over previous multi-target regression models is its great generality. Theoretically, the proposed MMR is highly generalized and encompasses some of existing models. By setting $S = I$ and $K = X^\top X$ in (9), the MMR can recover the fundamental multi-target ridge regression based on which many previous models were developed. The MMR can simultaneously encode inter-target correlations and disentangle linear/nonlinear input-output relationships by customizing kernels in one single framework. Moreover, it can also work with other convex loss functions, e.g., the $\varepsilon$-insensitive loss function [50] and accepts other regularization terms to satisfy desired properties [51]. Compared to the recent output kernel learning (OKL) [36], [52], the MMR employs a general low-rank regularization network without relying on specific assumptions, which allows to fully capture more complex, e.g., positive and negative, inter-target correlations rather than only a certain aspect, e.g., similarity of multiple targets [36]. Moreover, we do not assume that all tasks are correlated and allow the existence of outlier tasks [53], which further increases the generality.

### 3.4 Alternating Optimization

The obtained objective function (9) is non-trivial to solve simultaneously for $A$ and $S$ due to the non-convexity of the objective function. We derive a new alternating optimization algorithm to efficiently solve the objective function. Denote $J(A, S)$ as the objective function in (9), and we seek $A$ and $S$ alternately by solving $J(A, S)$ for one with the other fixed.

#### 3.4.1 Fix $S$ to optimize $A$

We calculate the gradients of the objective function with respect to $A$ as follows:

$$\frac{\partial J}{\partial A} = -\frac{1}{\mathcal{N}} S^\top (Y - SAK)K + \lambda AK. \tag{10}$$

Setting the derivatives to be **0** gives rise to

$$S^\top SAK + \lambda \mathcal{N} A = S^\top Y. \tag{11}$$

Multiplying $K^{-1}$ to both sides on the right leads to

$$S^\top SA + \lambda A \mathcal{N} K^{-1} = S^\top Y K^{-1}, \tag{12}$$

which is a standard Sylvester equation [54] and can be solved analytically in a closed form.

#### 3.4.2 Fix $A$ to optimize $S$

We propose a gradient based alterative optimization to solve for $S$, before which we provide the following proposition to calculate the derivative of $J$ w.r.t. $S$.

**Proposition 1.** *Assume that the singular value decomposition (SVD) of $S$ is*

$$S = U\Sigma V^\top, \tag{13}$$

*where $U$ and $V$ are unitary matrices and $\Sigma$ is the diagonal matrix with real numbers on the diagonal. Then the derivative of $||S||_*$ w.r.t. $S$ takes the form as follows:*

$$\frac{\partial ||S||_*}{\partial S} = U\Sigma^{-1}|\Sigma|V^\top \tag{14}$$

*where $\Sigma^{-1}$ is the Moore-Penrose pseudo-inverse of $\Sigma$.*

*Proof:* By the definition of the nuclear norm, we have

$$\begin{aligned} ||S||_* &= tr(\sqrt{S^\top S}) = tr(\sqrt{(U\Sigma V^\top)^\top (U\Sigma V^\top)}) \\ &= tr(\sqrt{V\Sigma U^\top U\Sigma V^\top}) = tr(\sqrt{V\Sigma^2 V^\top}) \end{aligned} \tag{15}$$

By the property of circularity of trace, we have

$$||S||_* = tr(\sqrt{V^\top V\Sigma^2}) = tr(\sqrt{\Sigma^2}) = tr(|\Sigma|) \tag{16}$$

where $|\Sigma|$ is the matrix of the element-absolute values of $\Sigma$. Therefore, the nuclear norm of $S$ can be also defined as the sum of the absolute values of the singular value decomposition of $S$. Although the absolute value function is not differentiable on every point in its domain, but we can find a subgradient.

$$\frac{\partial ||S||_*}{\partial S} = \frac{\partial tr(|\Sigma|)}{\partial S} = \frac{tr(\partial|\Sigma|)}{\partial S} \tag{17}$$

Since $\Sigma$ is diagonal, the subdifferential set of $|\Sigma|$ is:

$$\frac{\partial |\Sigma|}{\partial S} = |\Sigma|\Sigma^{-1}\frac{\partial \Sigma}{\partial S}. \tag{18}$$

By substituting (18) into (17), we obtain

$$\frac{\partial \|S\|_*}{\partial S} = \frac{tr(|\Sigma|\Sigma^{-1}\partial\Sigma)}{\partial S}. \tag{19}$$

From (13), we have $\partial S = \partial U\Sigma V^\top + U\partial\Sigma V^\top + U\Sigma\partial V^\top$, which gives rise to

$$U\partial\Sigma V^\top = \partial S - \partial U\Sigma V^\top - U\Sigma\partial V^\top. \tag{20}$$

Multiplying $U^\top$ on both sides of (20), we have

$$U^\top U\partial\Sigma V^\top V = U^\top \partial S V - U^\top \partial U\Sigma V^\top V - U^\top U\Sigma\partial V^\top V \tag{21}$$

which leads to

$$\partial\Sigma = U^\top \partial S V - U^\top \partial U\Sigma - \Sigma\partial V^\top V. \tag{22}$$

Note that $0 = \partial I = \partial(U^\top U) = \partial U^\top U + U^\top \partial U$, where $I$ is an identity matrix, and therefore $U^\top \partial U$ is an antisymmetric matrix. Since $\Sigma$ is a diagonal matrix, we have

$$\begin{aligned} tr(U^\top \partial U\Sigma) &= tr((U^\top \partial U\Sigma)^\top) = tr(\Sigma^\top \partial U^\top U) \\ &= -tr(\Sigma U^\top \partial U) = -tr(U^\top \partial U\Sigma) \end{aligned} \tag{23}$$

0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence



which indicates that $tr(U^\top \partial U \Sigma) = 0$. Similarly, we also have $tr(\Sigma \partial V^\top V) = 0$. Therefore, we achieve

$$\partial \Sigma = U^\top \partial S V \tag{24}$$

Substituting (24) into (19), we obtain

$$\begin{aligned} \frac{\partial \|S\|_*}{\partial S} &= \frac{tr(|\Sigma|\Sigma^{-1}\partial\Sigma)}{\partial S} = \frac{tr(|\Sigma|\Sigma^{-1}U^\top \partial S V)}{\partial S} \\ &= \frac{tr(V|\Sigma|\Sigma^{-1}U^\top \partial S)}{\partial S} \\ &= (V|\Sigma|\Sigma^{-1}U^\top)^\top \end{aligned} \tag{25}$$

which closes the proof. □

Proposition 1 associated with the rigorous proof provides a theoretical foundation, which can be directly used to solve a large while important family of optimization problems with trace norm minimization.

Based on the Proposition 1, we have the derivative of $J$ w.r.t $S$ as follows:

$$\begin{aligned} \frac{\partial J}{\partial S} =& -2\frac{1}{\mathcal{N}}(Y - SAK)(AK)^\top + \beta U\Sigma^{-1}|\Sigma|V^\top \\ &+ 2\gamma S \end{aligned} \tag{26}$$

where $U$, $\Sigma$ and $V$ are obtained by the SVD in (13).

Denote $\mathcal{G}(S)$ as the gradient w.r.t. $S$ in (26). Therefore, $S$ can be solved by an iterative optimization based on gradient descent.

$$S^{t+1} = S^t - \eta \mathcal{G}(S^t) \tag{27}$$

where $\eta$ is the step size also called learning rate, which can adaptively chosen by line search algorithms [55]. In each iteration, $S^{t+1}$ is calculated with the current $S^t$ associated with $U$, $\Sigma$ and $V$. Since the objective function $J(A, S)$ is convex with respect to $S$, it is guaranteed to find a global minimum of $S$.

Note that the size of $S$ depends only on the number $\mathcal{Q}$ of targets, which is usually much smaller than the dimensionality $d$ of inputs. Therefore, the complexity of the singular value decomposition (SVD) of the structure matrix $S$ involved in the calculation of the derivative of the nuclear norm is $\mathcal{O}(Q^3)$. This guarantees the efficiency of both the iterative algorithm to update $S$ and the alternating optimization algorithm (Algorithm 1).

#### 3.4.3 Proof of Convergence

The efficiency of the proposed MMR is ensured by the guaranteed convergence of the newly-derived alternating optimization algorithm. We provide theoretical analysis by rigorous proof of the convergence of the alternating optimization.

**Theorem 2.** *The objective function $J(A, S)$ in Sec. 3.4 is bounded from below and monotonically decreases with each optimization step for $A$ and $S$, and therefore it converges.*

*Proof:* Since $J(A, S)$ is the summation of norms, we have $J(A, S) \geq 0$ for any $A$ and $S$. Then $J(A, S)$ is bounded from below. Denote $A^{(t)}$ and $S^{(t)}$ as the $A$ and $S$ in the $t$-th iteration, respectively. For the $t$-th step, $A^{(t)}$ is computed by $A^{(t)} \leftarrow \arg\min_A J(A, S^{(t-1)})$. And we also have $J(A^{(t)}, S^{(t-1)}) \geq J(A^{(t)}, S^{(t)})$. In this way, we obtain the following inequality:

$$\begin{aligned} \cdots \geq J(A^{(t-1)}, S^{(t-1)}) \geq J(A^{(t)}, S^{(t-1)}) \\ \geq J(A^{(t)}, S^{(t)}) \geq \cdots . \end{aligned}$$

Therefore, $J(A^{(t)}, S^{(t)})$ is monotonically decreasing as $t \to +\infty$, which indicates that the objective function $J(A, S)$ converges according to the monotone convergence theorem. □

**Algorithm 1** Alternating Optimization

**Input:** Data matrices $X$ associated with corresponding targets $Y$, regularization parameters $\lambda$, $\beta$ and and $\gamma$.

**Output:** The regression coefficient matrix $A$ and the structure matrix $S$.

1: Randomly initialize $S \in \mathbb{R}^{\mathcal{Q}\times\mathcal{Q}}$ and set $i = 1$;
2: **repeat**
3: Calculate the matrix $A^{(i+1)}$ by solving the Sylvester equation in (12);
4: Calculate the $S^{(i+1)}$ using the iterative method based on gradient descent in (27);
5: $i \leftarrow i + 1$;
6: **until** Convergence.

#### 3.4.4 Complexity Analysis

The complexity of solving for $A$ arises from computing the inversion of the Gram matrix, which is $\mathcal{O}(N^3)$, where $N$ is the number of training samples. The total complexity of the alternating optimization algorithm is $\mathcal{O}(t_1 N^3)$ + $\mathcal{O}(t_2 Q^3 + t_1 Q N^2)$, where $t_1$ is the iterations of the whole alternating optimization and $t_2$ is the total iteration steps of updating $S$. The overall complexity is approximately $\mathcal{O}(N^3)$ due to the fact that $Q \ll N$, $t_1 \ll N$ and $t_2 \ll N$. Therefore, the complexity of the proposed MMR is roughly the same as regular kernel methods, e.g., KRR.

To deal with datasets of large scale or arriving sequentially, online learning [56], [57], e.g., sequential minimal optimization (SMO), and kernel approximation methods [58] can be employed. Kernel approximation has recently attracted increasing research efforts to speed up kernel methods. When using translation invariant kernels, a large family of most widely-used kernels, e.g., the radius basis function (RBF), the feature map induced by the kernel function can be approximated by random Fourier features [58] under Bochner's Theorem [59]. Stochastic gradient and back-propagation can be used to train the model to scale up with data of large scales or arriving sequentially.

0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence



# 4 Experiments and Results

We have conducted extensive experiments to evaluate the performance of the MMR on all the 18 real-world datasets and compared with state-of-the-art algorithms, and we have also investigated the convergence of the alternating optimization and showed the results on 2 representatives datasets with different amount of targets.

## 4.1 Datasets and Settings

The 18 real-world datasets are widely-used benchmarks for multi-target regression in [60], which cover a large range of multi-target prediction tasks. Inter-target correlations demonstrate diverse patterns on across different datasets, which poses great challenges for multi-target regression models. The statistics of these datasets are summarized in Table 1. We follow the strategies in [60] to process the datasets with missing values in inputs, which are replaced with sample means in the datasets.

We compare with existing representative multi-target regression models including multi-dimensional support vector regression (mSVR) [50], [61], output kernel learning (OKL) [36], adaptive k-cluster random forests (AKRF) [8], multi-task feature learning (MTFL) [7] and MROTS [13]. Note that MTRL [5] and FIRE [62] perform worse than MROTS and MORF, respectively [13] and are therefore not included for comparison. The methods in [60] including single task learning (STL), multi-object random forests (MORF), the corrected multi-target stacking (MSTC), ensemble of regressor chains (ERC) and random linear target combinations (RLC), which have shown great performance in [60], are also included for comprehensive comparison. We follow the evaluation settings in [60] to benchmark with other algorithms. Specifically, as shown in Table 1, we use two-fold cross validation (CV) for SCM1d/SCM20d, five-fold CV for RF1/RF2 and ten-fold CV for the rest of the datasets.

To directly benchmark with state-of-the-art algorithms, we measure the performance by the commonly-used Relative Root Mean Squared Error (RRMSE) defined as: $\mathrm{RRMSE} = \sqrt{\frac{\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in D_{test}}(\hat{\mathbf{y}}_i-\mathbf{y}_i)^2}{\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in D_{test}}(\hat{Y}-\mathbf{y}_i)^2}}$, where $(\mathbf{x}_i, \mathbf{y}_i)$ is the $i$-th sample $\mathbf{x}_i$ with ground truth target $\mathbf{y}_i$, $\hat{\mathbf{y}}_i$ is the prediction of $\mathbf{y}_i$ and $\hat{Y}$ is the average of the targets over the training set $D_{train}$. We take the average RRMSE (aRRMSE) across all the target variables within the test set $D_{test}$ as a single measurement. A lower aRRMSE indicates better performance. The parameters $\lambda$, $\beta$ and $\gamma$ are chosen by cross validation from a search grid of $10^{[-5:1:3]}$ on the training set by tuning one with the others fixed, which could also be selected by adaptive techniques explored in [63], [25]. We use the radial basis function (RBF) kernel for nonlinear regression.

TABLE 1: The statistics of the 18 datasets. $d$ is the dimension of inputs and $Q$ is the number of targets.

| Dataset | Samples | Input ($d$) | Target ($Q$) | #-Fold CV |
|---|---|---|---|---|
| EDM | 154 | 16 | 2 | 10 |
| SF1 | 323 | 10 | 3 | 10 |
| SF2 | 1066 | 10 | 3 | 10 |
| JURA | 359 | 15 | 3 | 10 |
| WQ | 1066 | 16 | 14 | 10 |
| ENB | 768 | 8 | 2 | 10 |
| SLUMP | 103 | 7 | 3 | 10 |
| ANDRO | 49 | 30 | 6 | 10 |
| OSCALES | 639 | 413 | 12 | 10 |
| SCPF | 1137 | 23 | 3 | 10 |
| ATP1d | 337 | 411 | 6 | 10 |
| ATP7d | 296 | 411 | 6 | 10 |
| OES97 | 334 | 263 | 16 | 10 |
| OES10 | 403 | 298 | 16 | 10 |
| RF1 | 9125 | 64 | 8 | 5 |
| RF2 | 9125 | 576 | 8 | 5 |
| SCM1d | 9803 | 280 | 16 | 2 |
| SCM20d | 8966 | 61 | 16 | 2 |

## 4.2 Results

### 4.2.1 Performance

The proposed MMR algorithm has achieved consistently high multi-target prediction performance on all 18 datasets and substantially outperforms state-of-the-art algorithms. The multi-target prediction results of the proposed MMR and the comparison with state-of-the-art algorithms that recently proposed are summarized in Table 2.

The proposed MMR substantially outperforms the best results from state-of-the-art algorithms on most of these 18 datasets except the ANDRO dataset with only 49 samples. The great effectiveness of the MMR has been validated for a broad range of multi-target regression tasks. The large improvement of the proposed MMR over the STL and mKRR with significant margins on the all the 18 datasets, which shows its effectiveness in modeling inter-target correlations. The STL and mKRR are regarded as the baseline methods that predict multiple targets independently without exploring the correlation among multiple targets. The MMR achieves The large improvement over these methods including MTSC, ERC, RLC, mSVR, AKRF, MROTS and OKL, in all of which the inter-target correlation is explored in certain way. This comparison results demonstrate the advantage of the proposed MEN in modeling inter-target correlations via robust low-rank learning by introducing a structure matrix.

### 4.2.2 Convergence

The fast convergence is very important for practical use of the MMR. The proposed alternating optimization in Algorithm 1 shows very quick convergence on all these 18 datasets. The algorithm converges within a few ($\leq 20$) iterations on all these datasets. We show the convergence of the alternative optimization on the two representative datasets, i.e., the EDM dataset with relatively small (2) targets and the SCM1d dataset

0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence



TABLE 2: The comparison with state-of-the-art algorithms on 18 real-world datasets in terms of aRRMSE (%).

| Dataset \ Method | STL | MTSC | ERC | RLC | MORF | mSVR | AKRF | MTFL | MROTS | OKL | mKRR | **MMR** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EDM | 74.2 | 74.0 | 74.1 | 73.5 | 73.4 | 73.7 | 74.0 | 85.1 | 81.2 | 74.1 | 83.3 | **71.6** |
| SF1 | 113.5 | 106.8 | 108.9 | 116.3 | 128.2 | 102.1 | 111.4 | 111.2 | 115.5 | 105.9 | 110.4 | **95.8** |
| SF2 | 114.9 | 105.5 | 108.8 | 122.8 | 142.5 | 104.3 | 113.5 | 112.7 | 120.1 | 100.4 | 116.6 | **98.4** |
| JURA | 58.9 | 59.1 | 59.0 | 59.6 | 59.7 | 61.1 | 61.8 | 60.8 | 62.5 | 59.9 | 63.3 | **58.2** |
| WQ | 90.8 | 90.9 | 90.6 | 90.2 | 89.9 | 89.9 | 91.8 | 96.2 | 91.3 | 89.1 | 92.0 | **88.9** |
| ENB | 11.7 | 12.1 | 11.4 | 12.0 | 12.1 | 22.0 | 23.4 | 31.6 | 25.7 | 13.8 | 26.3 | **11.1** |
| SLUMP | 68.8 | 69.5 | 68.9 | 69.0 | 69.4 | 71.1 | 72.9 | 68.1 | 77.8 | 69.9 | 78.9 | **58.7** |
| ANDRO | 60.2 | 57.9 | 56.7 | 57.0 | **51.0** | 62.7 | 62.3 | 80.3 | 63.5 | 55.3 | 63.9 | **52.7** |
| OSCALES | 74.8 | 72.6 | 71.3 | 74.1 | 75.3 | 77.8 | 77.5 | 168.2 | 80.0 | 71.8 | 79.9 | **70.9** |
| SCPF | 83.7 | 83.1 | 83.0 | 83.5 | 83.3 | 82.8 | 83.1 | 89.9 | 90.1 | 82.0 | 85.5 | **81.2** |
| ATP1d | 37.4 | 37.2 | 37.2 | 38.4 | 42.2 | 38.10 | 41.2 | 41.5 | 40.4 | 36.4 | 38.0 | **33.2** |
| ATP7d | 52.5 | 50.7 | 51.2 | 46.1 | 55.1 | 47.75 | 53.1 | 55.3 | 54.9 | 47.5 | 48.6 | **44.3** |
| OES97 | 52.5 | 52.4 | 52.4 | 52.3 | 54.9 | 55.7 | 58.1 | 81.8 | 60.5 | 53.5 | 58.7 | **49.7** |
| OES10 | 42.0 | 42.1 | 42.0 | 41.9 | 45.2 | 44.7 | 44.6 | 53.2 | 55.8 | 43.2 | 48.9 | **40.3** |
| RF1 | 9.7 | 9.4 | 9.1 | 12.1 | 12.3 | 10.9 | 11.4 | 98.3 | 15.4 | 11.2 | 17.9 | **8.9** |
| RF2 | 10.2 | 9.7 | **9.5** | 13.0 | 14.8 | 14.4 | 15.7 | 110.3 | 19.8 | 11.8 | 15.9 | **9.5** |
| SCM1d | 34.8 | 33.6 | 33.0 | 34.5 | 35.2 | 36.7 | 36.8 | 43.7 | 44.9 | 34.2 | 37.1 | **31.8** |
| SCM20d | 47.5 | 41.3 | 39.4 | 44.3 | 48.2 | 49.3 | 65.5 | 64.3 | 45.6 | 44.3 | 49.8 | **38.9** |

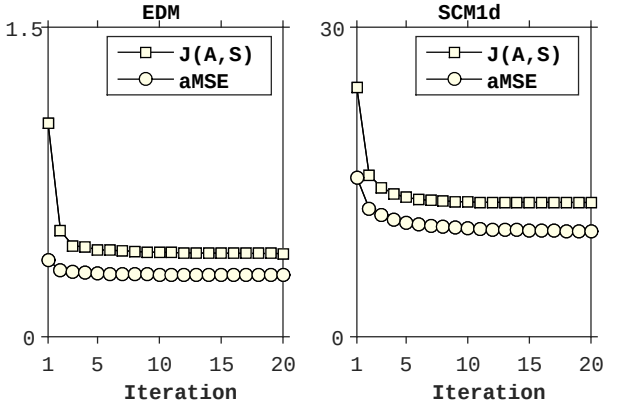


Fig. 2: The convergence of the proposed alternating optimization algorithm on the two representative datasets, i.e., EDM and SCM1d with 2 and 16 targets, respectively. $J(A, S)$ is the value of the objective function and aMSE is the average mean square error.

with relatively larger (16) targets. The convergence with respect to the iteration steps is plotted in Fig. 2. Both the objective function value and the average mean square error (aMSE) decrease monotonously with alternation steps. Although we show the first 20 steps, the algorithm can converge within only 10 iterations on the EDM dataset and within 15 iterations on the SCM1d dataset. The consistently quick convergence shows the great efficiency of the alternative optimization and guarantees the practical implementation of the MMR.

## 5 Conclusion

We have presented a multi-layer multi-target regression (MMR) framework that enables simultaneously modeling inter-target correlations and nonlinear input-output relationships. The MMR introduces a latent space to explicitly encode inter-target correlations in a structure matrix which is learned by robust low-rank learning via matrix elastic nets (MEN) from data without relying any specific assumptions; the MMR is flexible and can seamlessly work in conjunction with the kernel trick, which enables to handle highly complex nonlinear relationships between high-dimensional inputs and multiple targets; the MMR can be solved efficiently by a alternating optimization algorithm with theoretically guaranteed convergence. The MMR combines the strengths of kernel methods for nonlinear feature learning and the structural advantage of multi-layer architectures to capture inter-target correlations. More importantly, it offers a new multi-layer learning paradigm for multi-target regression, which is endowed with high generality, flexibility and expressive ability. Extensive experiments have been conducted on 18 real-world datasets, which validates the great effectiveness and generality of the MMR for diverse multivariate prediction.

## References

[1] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
[2] Y. Wang, D. Wipf, Q. Ling, W. Chen, and I. Wassell, "Multi-task learning for subspace segmentation," in *ICML*, 2015, pp. 1209–1217.
[3] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe, "A multi-task learning framework for head pose estimation under target motion." *TPAMI*, 2015.
[4] X. Zhen, Z. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Multi-scale deep networks and regression forests for direct bi-ventricular volume estimation," *Medical Image Analysis*, 2015.
[5] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," *UAI*, 2010.
[6] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *JMLR*, vol. 6, pp. 1817–1853, 2005.
[7] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *NIPS*, 2006, pp. 41–48.
[8] K. Hara and R. Chellappa, "Growing regression forests by classification: Applications to object pose estimation," in *ECCV*, 2014, pp. 552–567.

0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2688363, IEEE Transactions on Pattern Analysis and Machine Intelligence




[9] M. Long and J. Wang, "Learning transferable features with deep adaptation networks," *ICML*, 2015.

[10] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, 2015.

[11] A. J. Rothman, E. Levina, and J. Zhu, "Sparse multivariate regression with covariance estimation," *JCGS*, vol. 19, no. 4, pp. 947–962, 2010.

[12] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194–200, 2011.

[13] P. Rai, A. Kumar, and H. Daume, "Simultaneously leveraging output and task structures for multiple-output regression," in *NIPS*, 2012, pp. 3185–3193.

[14] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[15] L. Jacob, J.-p. Vert, and F. R. Bach, "Clustered multi-task learning: A convex formulation," in *NIPS*, 2009, pp. 745–752.

[16] L. Han and Y. Zhang, "Learning tree structure in multi-task learning," in *SIGKDD*, 2015, pp. 397–406.

[17] H. Liu, L. Wang, and T. Zhao, "Calibrated multivariate regression with application to neural semantic basis discovery," *JMLR*, vol. 16, pp. 1579–1606, 2015.

[18] F. Dinuzzo and B. Schölkopf, "The representer theorem for Hilbert spaces: a necessary and sufficient condition," in *NIPS*, 2012, pp. 189–196.

[19] K. Yu, V. Tresp, and A. Schwaighofer, "Learning gaussian processes from multiple tasks," in *ICML*, 2005, pp. 1012–1019.

[20] H. Daumé III, "Bayesian multitask learning with latent hierarchies," in *UAI*, 2009, pp. 135–142.

[21] A. Kumar and H. Daume, "Learning task grouping and overlap in multi-task learning," in *ICML*, 2012, pp. 1383–1390.

[22] M. Álvarez, L. Rosasco, and N. Lawrence, *Kernels for Vector-Valued Functions: A Review*, ser. Foundations and Trends in Machine Learning, 2012.

[23] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," in *JMLR*, 2005, pp. 615–637.

[24] A. Bargi, M. Piccardi, and Z. Ghahramani, "A non-parametric conditional factor regression model for multi-dimensional input and response," in *AISTATS*, 2014.

[25] H. Li, N. Chen, and L. Li, "Error analysis for matrix elastic-net regularization algorithms," *TNNLS*, vol. 23, no. 5, pp. 737–748, 2012.

[26] A. G. Wilson, D. A. Knowles, and Z. Ghahramani, "Gaussian process regression networks," *ICML*, 2012.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[28] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep kernel learning," in *AISTATS*, 2016.

[29] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization," in *UAI*, 2009, pp. 339–348.

[30] X. Wang, J. Bi, S. Yu, and J. Sun, "On multiplicative multitask feature learning," in *NIPS*, 2014, pp. 2411–2419.

[31] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," *TKDD*, vol. 5, no. 4, p. 22, 2012.

[32] C. Ciliberto, Y. Mroueh, T. Poggio, and L. Rosasco, "Convex learning of multiple tasks and their structure," in *ICML*, 2015, pp. 1548–1557.

[33] Q. Zhou and Q. Zhao, "Flexible clustered multi-task learning by learning representative tasks." *TPAMI*, vol. 38, no. 2, pp. 266–278, 2016.

[34] L. Xu, Z. Wang, Z. Shen, Y. Wang, and E. Chen, "Learning low-rank label correlations for multi-label classification with missing labels," in *ICDM*, 2014, pp. 1067–1072.

[35] H. Liu, L. Wang, and T. Zhao, "Multivariate regression with calibration," in *NIPS*, 2014, pp. 127–135.

[36] F. Dinuzzo, C. S. Ong, G. Pillonetto, and P. V. Gehler, "Learning output kernels with block coordinate descent," in *ICML*, 2011, pp. 49–56.

[37] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[38] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[39] X. Zhong, L. Xu, Y. Li, Z. Liu, and E. Chen, "A nonconvex relaxation approach for rank minimization problems." in *AAAI*, 2015, pp. 1980–1987.

[40] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[41] F. Nie, H. Huang, and C. Ding, "Low-rank matrix recovery via efficient schatten p-norm minimization," in *AAAI*, 2012, pp. 655–661.

[42] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *JRSS*, vol. 67, no. 2, pp. 301–320, 2005.

[43] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[44] J. Gillberg, P. Marttinen, M. Pirinen, A. J. Kangas, P. Soininen, M. Ali, A. S. Havulinna, M.-R. M.-R. Järvelin, M. Ala-Korpela, and S. Kaski, "Multiple output regression with latent noise," *arXiv preprint arXiv:1410.7365*, 2014.

[45] P. Gong, J. Zhou, W. Fan, and J. Ye, "Efficient multi-task feature learning with calibration," in *SIGKDD*, 2014, pp. 761–770.

[46] G. S. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *The Annals of Mathematical Statistics*, vol. 41, no. 2, pp. 495–502, 1970.

[47] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *NIPS*, 2010, pp. 1813–1821.

[48] S. Zheng, X. Cai, C. Ding, F. Nie, and H. Huang, "A closed form solution to multi-view low-rank regression," in *AAAI*, 2015.

[49] B. Rakitsch, C. Lippert, K. Borgwardt, and O. Stegle, "It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals," in *NIPS*, 2013, pp. 1466–1474.

[50] M. Sánchez-Fernández, M. de Prado-Cumplido, J. Arenas-García, and F. Pérez-Cruz, "SVM multiregression for nonlinear channel estimation in multiple-input multiple-output systems," *TSP*, vol. 52, no. 8, pp. 2298–2307, 2004.

[51] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[52] F. Dinuzzo, "Learning output kernels for multi-task problems," *Neurocomputing*, vol. 118, pp. 119–126, 2013.

[53] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *SIGKDD*, 2012, pp. 895–903.

[54] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[55] L. Armijo, "Minimization of functions having lipschitz continuous first partial derivatives," *Pacific Journal of mathematics*, vol. 16, no. 1, pp. 1–3, 1966.

[56] J. C. Platt, "12 fast training of support vector machines using sequential minimal optimization," *Advances in kernel methods*, pp. 185–208, 1999.

[57] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE TSP*, vol. 52, no. 8, pp. 2165–2176, 2004.

[58] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *NIPS*, 2007, pp. 1177–1184.

[59] K. Yano, "On harmonic and killing vector fields," *Annals of Mathematics*, pp. 38–45, 1952.

[60] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: treating targets as inputs," *Machine Learning*, vol. 104, no. 1, pp. 55–98, 2016.

[61] D. Tuia, J. Verrelst, L. Alonso, F. Pérez-Cruz, and G. Camps-Valls, "Multioutput support vector regression for remote sensing biophysical parameter estimation," *TGRSL*, vol. 8, no. 4, pp. 804–808, 2011.

[62] T. Aho, B. Ženko, S. Džeroski, and T. Elomaa, "Multi-target regression with rule ensembles," *JMLR*, vol. 13, no. 1, pp. 2367–2407, 2012.

[63] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *Annals of statistics*, vol. 37, no. 4, p. 1733, 2009.


0162-8828 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.