Multitarget Sparse Latent Regression

Xiantong Zhen, Mengyang Yu, *Student Member, IEEE*, Feng Zheng, Ilanit Ben Nachum, Mousumi Bhaduri, David Laidley, and Shuo Li*

***Abstract*— Multitarget regression has recently generated intensive popularity due to its ability to simultaneously solve multiple regression tasks with improved performance, while great challenges stem from jointly exploring inter-target correlations and input-output relationships. In this paper, we propose multitarget sparse latent regression (MSLR) to simultaneously model intrinsic intertarget correlations and complex nonlinear input–output relationships in one single framework. By deploying a structure matrix, the MSLR accomplishes a latent variable model which is able to explicitly encode intertarget correlations via $\ell_{2,1}$-norm-based sparse learning; the MSLR naturally admits a representer theorem for kernel extension, which enables it to flexibly handle highly complex nonlinear input–output relationships; the MSLR can be solved efficiently by an alternating optimization algorithm with guaranteed convergence, which ensures efficient multitarget regression. Extensive experimental evaluation on both synthetic data and six greatly diverse real-world data sets shows that the proposed MSLR consistently outperforms the state-of-the-art algorithms, which demonstrates its great effectiveness for multivariate prediction.**

***Index Terms*— $\ell_{2,1}$-norm, latent variable models, multitarget regression, sparse learning,**

I. Introduction

MULTITARGET regression [1] has recently generated increasing popularity in the machine learning community due to its great capability of predicting multiple outputs simultaneously with improved generalization performance [2], [3]. The core of multitarget regression is to model the intrinsic intertarget correlation, which can largely improve parameter estimation by sharing knowledge across correlated outputs for more accurate multitarget prediction [4]. However, intertarget correlations vary greatly according to different applications [5], which requires learning automatically from data to cater the application. Moreover, multiple outputs that represent higher level concepts generate hugely complex relationships with the high-dimensional inputs [6], which demands powerful nonlinear regression models. The major challenges of multitarget regression lie in jointly modeling both intertarget correlations and complex input–output relationships. However, these two challenges have not yet been well addressed simultaneously in one single framework.

Manuscript received February 13, 2016; revised May 31, 2016 and September 11, 2016; accepted December 31, 2016. *(Corresponding author: Shuo Li.)*

X. Zhen, I. B. Nachum, and S. Li are with the Department of Medical Biophysics, University of Western Ontario, London, ON, Canada.

M. Yu is with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K.

F. Zheng is with the Department of Electrical and Electronic Engineering, The University of Sheffield, Sheffield, U.K.

M. Bhaduri and D. Laidley are with the London Health Sciences Centre, London, ON, Canada.

Digital Object Identifier 10.1109/TNNLS.2017.2651068

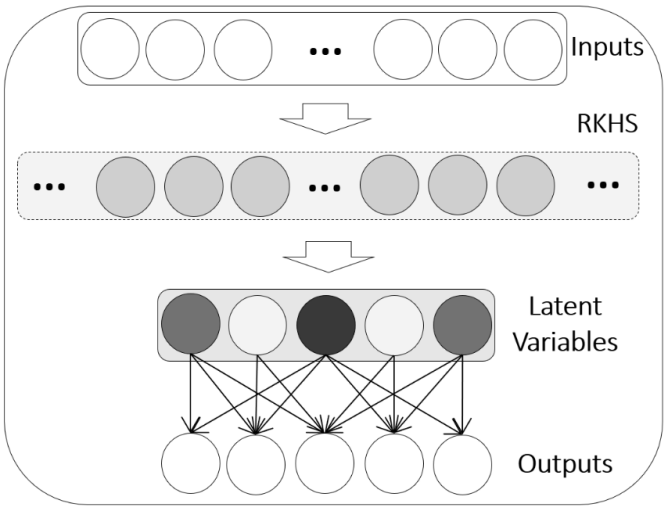


Fig. 1. Illustration of MSLR.

Previous multitarget regression models were focused on linear regression models [7]–[9] or specifically developed under particular assumptions with strong prior knowledge to facilitate the correlation modeling [5], [10]. However, these linear models suffer from the limited ability to handle nonlinear relationships between high-dimensional inputs and multiple outputs [6]; it is nontrivial to extend these linear models for nonlinear regression due to the nonconvexity of sparsity constraints or loss functions [9], [11]. Those particular assumptions, e.g., regression task parameters share a common prior [12], [13], or share a linear subspace [14], [15], can be too restrictive and would not necessarily hold or be shared by different applications in practice [16].

In this paper, we propose multitarget sparse latent regression (MSLR) to simultaneously model intertarget correlations and highly complex nonlinear input–output relationships in one general framework. In contrast to existing methods, the MSLR incorporates a latent space to explicitly encode intertarget correlations without relying on specific assumptions while being able to disentangle nonlinear input–output relationships by working with the kernel trick.

The learning architecture of the proposed MSLR is shown in Fig. 1. The inputs are embedded into an infinite-dimensional reproducing kernel Hilbert space (RKHS) induced by some nonlinear kernel, which serves to disentangle complex input–output relationships. The latent variables in the latent space obtained via a linear representer theorem [17] extract higher level concepts to build a common representation for multiple regression outputs. The intertarget correlation is explicitly modeled by a structure matrix which is learned from data via

the $\ell_{2,1}$-norm-based [18] sparse learning in a data-driven way to cater different applications.

In contrast to existing multitarget regression models, the proposed MSLR offers multiple attractive merits.

1) By incorporating a latent space, the MSLR accomplishes a new multitarget regression framework of a latent variant model, which enables simultaneously modeling intertarget correlations and input–output relationships.
2) By deploying a structure matrix, the MSLR provides a sparse learning framework to explicitly model intertarget correlations, which enables it to learn the correlations automatically from data without relying on any specific prior assumptions.
3) By working seamlessly with the kernel trick, the MSLR can disentangle the nonlinear relationship between inputs and outputs, which enables it to handle more complex multitarget regression tasks.

The effectiveness and the generality of the proposed MSLR have been validated by extensive evaluation on both synthetic data and six challenging real-word data sets for diverse multivariate prediction tasks. The MSLR consistently achieves high performance for all tasks and substantially outperforms the state-of-the art multitarget regression algorithms.

Notations: The $\ell_{2,1}$-norm $||\cdot||_{2,1}$ [18] is the sum of the Euclidean norms of the rows of the matrix, and is defined as

$$||A||_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} A_{i,j}^2} = \sum_{i=1}^{n} ||\mathbf{a}^i||_2$$

where $A_{i,j}$ is the element in the ith row and the jth column of A and $\mathbf{a}^i$ is the ith row of A. The $\ell_{2,1}$-norm is different from the ℓ_1-norm which is simply the maximum absolute column sum of the matrix, and the $\ell_{2,1}$-norm is a special case of the $\ell_{p,q}$-norm with $p = 2$ and $q = 1$, which is defined as

$$||A||_{p,q} = \left[\sum_{i=1}^{n} \left(\sum_{j=1}^{m} |A_{i,j}|^p \right)^{q/p} \right]^{1/q}. \quad (1)$$

$||A||_F^2$ is the Frobenius norm of matrix A, and can be computed by $tr(A^\top A)$, where $tr(\cdot)$ is the trace operator for a matrix.

II. Related Work

Previous models were mainly focused on imposing regularization terms on the regression weights to explore output relationships, which, however, would not be able to effectively handle highly complex input-output relationships and intertarget correlations simultaneously. They were built on particular aspects, e.g., simply learning features for multiple tasks [5], [19]–[21] or solely exploring specific task structures [4], [22]–[24], some of which were developed only for classification tasks [25], [26]. We briefly review the representative work recently proposed in the literature and refer to [27] for an up-to-date comprehensive survey.

Rothman *et al.* [7] proposed a multivariate regression model with covariance estimation (MRCE), in which a procedure is developed for constructing a sparse estimator of a multivariate regression coefficient matrix that accounts for correlation of the response variables. However, the MRCE does not leverage the learned output structure to share similar input variables among related outputs [28]. Moreover, it is a linear regression model with limited ability to handle nonlinear regression tasks.

Zhang and Yeung [16] proposed a convex formulation for multitask relationship learning (MTRL), which models the relationships between tasks in a nonparametric manner based on the assumption that all tasks are close to each other by measuring the Frobenius norms of their differences. To facilitate task relationship modeling, the MTRL is derived by placing prior assumptions of multivariate normal distributions on both multiple outputs and regression parameters.

Rai *et al.* [8] proposed multitarget regression with output and task structures (MROTS), which jointly explores the covariance structure of latent model parameters and the conditional covariance structure of multiple outputs. MROTS outperforms both MTRL and MRCE and theoretically generalizes them as its special cases [7], [16]. However, similar to the MRCE [7], the MROTS does not provide any mechanism for nonlinear regression.

By introducing a matrix ℓ_1-norm based inverse-covariance regularization, Sohn and Kim [28] proposed joint estimation of structured sparsity and output structure for multitarget regression. Based also on linear regression as in [7] and [8], the method assumes that the output structure of multiple outputs can be represented as a graph. Recently, the output kernel learning (OKL) was developed for vector-valued functions to explore intertarget correlations for multiple task learning [29]–[31], which would not fully capture intertarget correlations by simply learning a semidefinite similarity matrix, i.e., the output kernel, of multiple outputs.

Ensemble algorithms have also been explored for multitarget learning. Aho *et al.* [32] introduced the fitted rule ensembles (FIREs) algorithm to improve multitarget regression by adding simple linear functions to the ensemble. However, the FIRE algorithm performs slightly worse than the multiobjective random forests (MORF) [33]. Tsoumakas *et al.* [34] presented an ensemble method which constructs new target variables by random linear combination (RLC) of existing outputs, which is also heuristically derived from the counterpart in multilabel classification.

Multitarget stacking (MTS) [35] and ensemble of regressor chains (ERC) [36] are introduced in [37] by transferring from multilabel classification. The basic idea is to decompose multitarget problem into multiple single-target problems, in which individual outputs are predicted by treating the rest as additional input variables. However, the fact that the training and testing data should be identically and independently distributed is ignored. To compensate this, the modified versions: MTSC and ERCC like MTS and ERC are developed, which, however, do not provide an explicit formula to model intertarget correlations.

Recently, Bargi *et al.* [38] proposed a nonparametric conditional factor regression (NCFR) model for multitarget regression to enhance linear regression. The NCFR introduces low-dimensional latent factors with an Indian Buffet Process prior to improve the model by decoupling inputs and outputs into separate noise models. However, its capability of non-

linear regression is also limited due to the nature of linear regression.

Rather than assuming all tasks to be relevant, the clustered multitask learning (CMTL) [39] assumes that all tasks can be clustered into disjoint groups. Very recently, to improve the performance of the CMTL, Zhou and Zhao [24] presented an enhanced CMTL called flexible CMTL, which learns the cluster structure by identifying representative tasks. However, the assumption of the existence of representative tasks would not necessarily hold for due to the diversity of different applications.

In addition, multitarget regression has recently started find successful applications in both computer vision [6], [40], [41] and medical image analysis [42], [43], showing great advantages over conventional approaches. By casting into multitarget regression, traditional tasks can solved more efficiently in more compact formulations [43].

Unlike these existing methods, the proposed MSLR incorporates a latent space associated with a structure matrix, which achieves a general framework that enables explicitly modeling intertarget correlations while jointly handling complex input–output relationships. The proposed MSLR has been extensively evaluated on both the synthetic and six diverse, real-world benchmark data sets released very recently [37], showing great effectiveness and generality for diverse multitarget regression tasks. The MSLR has also shown great effectiveness in medical image analysis for shape regression [43] and cardiac four chamber volume estimation [44].

III. Multitarget Sparse Latent Regression

The proposed MSLR incorporates a latent space from which a structure matrix is learned to explicitly model inter-target correlations via the $\ell_{2,1}$-norm-based sparse learning without unnecessary assumptions (Section III-B); the MSLR provides a natural formulation to work in conjunction with the kernel trick to effectively tackle nonlinear input–output relationships (Section III-C); the proposed MSLR can be solved efficiently with a newly derived alternating optimization algorithm with guaranteed convergence (Section III-D).

A. Multitarget Regression

Multitarget regression is to predict multiple continuous variables from an input vector. We consider the fundamental linear multitarget regression model

$$\mathbf{y} = W\mathbf{x} + \mathbf{b} \tag{2}$$

where $\mathbf{y} = [y_1, \ldots, y_i, \ldots, y_Q]^\top \in \mathbb{R}^Q$ are the multivariate outputs, $\mathbf{x} \in \mathbb{R}^d$ is the input, $W = [\mathbf{w}_1, \ldots, \mathbf{w}_i, \ldots, \mathbf{w}_Q]^\top \in \mathbb{R}^{Q\times d}$ is the model parameter, i.e., the regression coefficient or weight matrix, each $\mathbf{w}_i \in \mathbb{R}^d$ is the predictor for y_i, $\mathbf{b} \in \mathbb{R}^Q$ is the bias, and d and Q are dimensions of input and output spaces, respectively.

Given the training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, one can solve for W by the following objective:

$$W^* = \arg\min_W L(W) + \lambda\Omega(W) \tag{3}$$

where $L(W)$ is the empirical loss function which is assumed to be convex, $\Omega(W)$ is the regularization term that usually penalizes the complexity of the model parameter W to avoid overfitting, and $\lambda > 0$ is the regularization parameter.

Note that our method can work with different loss functions $L(W)$ and regularization terms $\Omega(W)$ to favor desirable properties of solutions for different applications. Without loss of generality, we start deriving our MSLR with the least square loss function and ℓ_2 regularization due to their important and fundamental roles played in regression models, e.g., the multitarget ridge regression

$$W^* = \arg\min_W \frac{1}{N}||Y - WX - B||_F^2 + \lambda||W||_F^2 \tag{4}$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$ and $B = [\ldots, \mathbf{b}, \ldots] \in \mathbb{R}^{Q\times N}$. Although the model in (4) can predict multiple outputs simultaneously, it does not provide any learning mechanism to explore the intertarget correlation, which, however, fails to fulfill the advantage of multitask learning to leverage the shared knowledge across correlated outputs for improved performance. Previous work has focused on imposing constraints, e.g., sparsity, on the regression matrix W to explore intertarget correlations, which, however, would compromise the flexibility and expressive ability of the model.

B. Sparse Latent Regression

We propose introducing a latent space with latent variables $\mathbf{z}$ based on which a structure matrix U is employed to explicitly model intertarget correlations. To effectively capture the intrinsic correlations, we propose imposing an $\ell_{2,1}$-norm-based sparsity constraint on S, which achieves sparse learning

$$\min_{W,U} \tfrac{1}{N}||Y - UZ||_F^2 + \lambda||W||_F^2 + \beta||U^\top||_{2,1} \tag{5}$$

where $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\} = WX + B \in \mathbb{R}^{Q\times N}$ and $U \in \mathbb{R}^{Q\times Q}$ is the structure matrix; the $\ell_{2,1}$-norm constraint on U in (5) encourages to learn an U of column sparsity, which enables to capture the underlying structure of intertarget correlations by sharing subsets of latent variables (features) among correlated outputs; β is the regularization parameter to control the column sparsity of U. The structure matrix U is learned in a data-driven way without relying on any specific assumptions, which allows to automatically infer intrinsic intertarget correlations from data to cater different applications of great diversity.

Thanks to the $\ell_{2,1}$-norm based sparse learning of U, predictors of correlated outputs are encouraged to share similar parameter sparsity patterns to capture a common set of features, i.e., latent variables in the latent space [20]. Therefore, knowledge is shared by correlated outputs and the performance of multiple predictors can be significantly improved, which enables more accurate multitarget prediction. In contrast to existing multitarget regression models, imposing the sparsity constraint on the structure matrix U rather than on the regression coefficient matrix W can dramatically enhance model flexibility that enables kernel extension for nonlinear regression. Moreover, due to the introduced latent space,

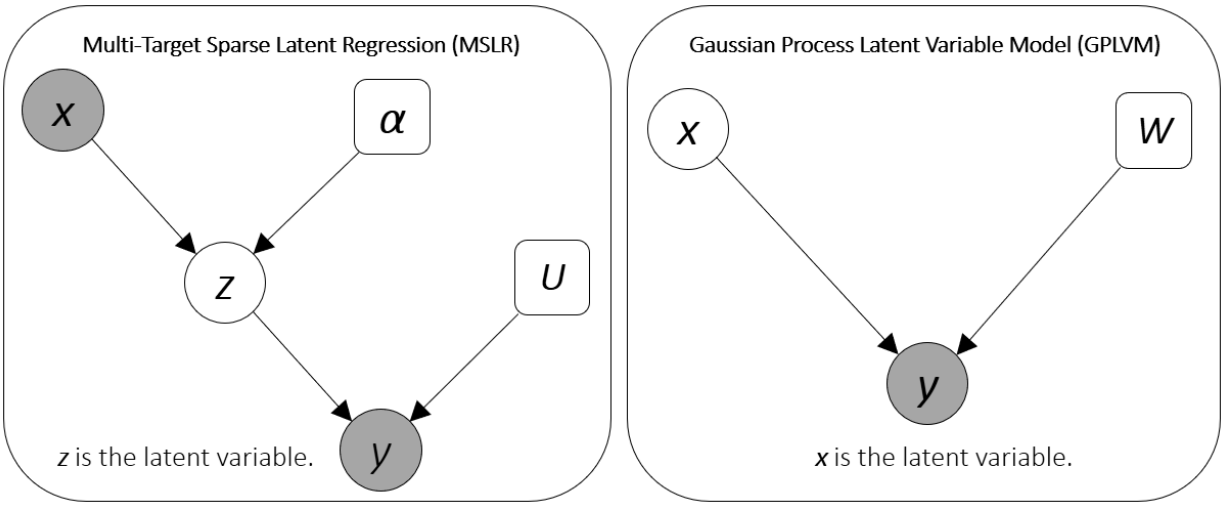


Fig. 2. Graphical illustration of MSLR versus GPLVM from the perspective of latent variable models. In the MSLR on the left side, $\mathbf{z}$ is the induced latent variable, which is obtained via linear transformation $\boldsymbol{\alpha}$ from the input $\mathbf{x}$; the multivariate output $\mathbf{y}$ is obtained via a structure matrix U by $\ell_{2,1}$-norm-based sparse learning. $\boldsymbol{\alpha}$ and U are the model parameters to be jointly learned in one single framework.

the MSLR accomplishes a new latent variable model with more expressive capability.

We highlight that due to the incorporation of the structure matrix U, the proposed MSLR accomplishes a new latent variable model with multiple favorable properties.

1) The latent space enables effectively dealing with inputs and outputs from distinctive distributions by decoupling them with the regression coefficient W and the structure matrix U [38], [45].
2) The structure matrix enables explicitly encoding intrinsic intertarget correlations, which is learned from data to cater different applications without relying on any specific assumptions on intertarget correlations.

The way that the proposed MSLR introduces latent variables is fundamentally different from existing latent variable models, e.g., generative graphical models (PGM) [46]. In contrast to PGMs, latent variables in the MSLR are incorporated in a nonprobabilistic, discriminant way, which not only avoids expensive, even intractable computation of partition functions and but also allows efficient inference by supervised learning. The MSLR also distinguishes from previous representative, important latent variable models, e.g., the relevance vector machine (RVM) [47] and the Gaussian process latent variable model (GPLVM) [48] both of which are probabilistic models. In the RVM, model parameters/weights are treated as latent variables by placing a Gaussian prior governed by a set of hyperparameters and are inferred iteratively by the expectation maximization optimization; in the GPLVM, the relationship between latent variables and input data is modeled by a Gaussian process, and latent variables are representations of input data, which are inferred by maximum likelihood estimation and do not necessarily have unique solutions; our MSLR is a discriminant learning model and the latent variables are lower dimensional, but high-level representations of input data, which can be efficiently optimized by supervised learning. The graphical illustration of the contrast between the GPLVM and the proposed MLSR is shown in Fig. 2.

C. Kernelization

The objective function in (5) remains a linear multitarget regression, being less able to effectively tackle complex input–output relationships, which are usually highly nonlinear. However, although the objective function in (5) is not jointly convex with respect to W and U, we will show in Theorem 1 that given a fixed U, (5) admits a linear representer theorem [17] with respect to W. Based on the linear representer theorem, kernel extension can be conveniently developed to achieve kernel multitarget regression.

Theorem 1: Assume we have the objective function in (5) defined over a Hilbert space $\mathcal{H}$. Given any fixed matrix U, if (5) has a minimizer with respect to W, it admits a linear representer theorem, that is

$$W = \boldsymbol{\alpha} X^{\top} \tag{6}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{Q \times N}$ is the coefficient matrix.

Remark 1: Theorem 1 provides important theoretical guarantees for kernel extension to achieve nonlinear regression. By specifying the kernel function, the MSLR can flexibly deal with linear/nonlinear input-output relationships. In contrast to previous methods imposing constraints directly on the regression matrix W, which could lose the nice property for kernel extension, the MSLR provides a natural formula to work in conjunction with kernels due to the introduction of the structure matrix U associated with the latent space. It is therefore allowed to simultaneously handle nonlinear input–output relationships and intrinsic intertarget correlations in one single framework. More importantly, Theorem 1 indicates that we can devise an alternating optimization algorithm to jointly solve for W and U, which is crucial to its practical applications. Although the proof of Theorem 1 is straightforward, it is theoretically of great importance and will offer a theoretical foundation to design new models, and we provide the rigorous proof as follows for theoretical completeness.

Proof: Since U is given and fixed, the third term in (5) is constant and can be dropped. Therefore, we denote

$$F_U(W) = \frac{1}{N}||Y - U(WX - B)||_F^2 + \lambda||W||_F^2. \quad (7)$$

The remaining two terms are quadratic, and therefore, (7) is convex with respect to W and has a minimizer. The regularization term is strictly monotonically increasing real-valued function, and the loss function is bounded and point wise [11]. According to the representer theorem [17], (7) admits a linear representer theorem, which takes the form of (6). □

The linear representer theorem will show great power when $\mathcal{H}$ is an RKHS, which allows to implicitly map input features into a high, even infinite dimensional space. By applying Theorem 1, we are now able to derive the kernel version of multitarget sparse latent regression in the RKHS to handle nonlinear input–output relationships. The objective function (5) can be rewritten in terms of traces as follows:

$$\min_{W,U} \frac{1}{N} tr((Y - U(WX - B))^\top (Y - U(WX - B))) + \lambda tr(W^\top W) + \beta||U^\top||_{2,1}. \quad (8)$$

Assume that $\mathbf{x}_i$ is mapped to $\phi(\mathbf{x}_i)$ in some RKHS of high, even infinite dimensionality, where $\phi(\cdot)$ denotes the feature map of $\mathbf{x}_i$. The associated kernel function is $k(\cdot, \cdot)$, that is, $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$. By applying the linear representer theorem in Theorem 1, the regression matrix W can be represented by

$$W = \boldsymbol{\alpha}\Phi(X)^\top \quad (9)$$

where $\Phi(X) = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_i), \ldots, \phi(\mathbf{x}_N)]$ and $\boldsymbol{\alpha} \in \mathbb{R}^{Q \times N}$. Plugging (9) into (8) gives rise to the following objective function:

$$\min_{\boldsymbol{\alpha},U} \frac{1}{N} tr((Y - U\boldsymbol{\alpha}\Phi(X)^\top \Phi(X))^\top (Y - U\boldsymbol{\alpha}\Phi(X)^\top \Phi(X))) + \lambda tr((\boldsymbol{\alpha}\Phi(X)^\top)^\top (\boldsymbol{\alpha}\Phi(X)^\top)) + \beta||U^\top||_{2,1}. \quad (10)$$

Define $K = \Phi(X)^\top \Phi(X)$ to be the kernel matrix in the RKHS space. We establish the kernel MSLR

$$\min_{\boldsymbol{\alpha},U} \frac{1}{N} tr((Y - U\boldsymbol{\alpha}K)^\top (Y - U\boldsymbol{\alpha}K)) + \lambda tr(\boldsymbol{\alpha}K\boldsymbol{\alpha}^\top) + \beta||U^\top||_{2,1}. \quad (11)$$

Note that the bias $\mathbf{b}$ is omitted, since it can be absorbed into W by adding additional dimension into the inputs $\mathbf{x}$ [49], [50]. The latent space obtained by the linear transformation $\boldsymbol{\alpha}$ via the representer Theorem from the KRHS induced by a nonlinear kernel contains higher level concepts [51] which not only fill the semantic gap between low-level inputs and high-level outputs but also facilitate efficient linear sparse learning of U to capture intertarget correlations. Once the model $(U, \boldsymbol{\alpha})$ in (11) is learned, the multiple outputs of a new input $\mathbf{x}_t$ can be predicted by

$$\hat{\mathbf{y}}_t = U\boldsymbol{\alpha}K_t \quad (12)$$

where $K_t = \Phi(X)^\top \phi(\mathbf{x}_t)$ and $\phi(\mathbf{x}_t)$ is the feature map of $\mathbf{x}_t$.

The MSLR provides a new multitarget regression framework of a latent variable model, which does not rely on specific loss functions and accepts different regularization terms. The MSLR can work with other convex loss functions, e.g., the ε-insensitive loss function in the multidimensional support vector regression (mSVR) [49] and regularization terms, which allows to cater different applications. The MSLR can flexibly deal with linear and nonlinear input–output relationships by prescribing the kernel in (11). It is easy to incorporate the prior knowledge of intertarget correlations by imposing additional constraints on the structure matrix U to further enhance the model. The generality of the proposed MSLR is demonstrated by Theorem 2, in which the MSLR recovers fundamental multitarget ridge regression via simple linear algebra by specifying the structure matrix and kernels.

Theorem 2: If the structure matrix U is set to be an identity matrix and the RKHS kernel K is a linear kernel, i.e., $K = X^\top X$, the solutions of (4) and (11) coincide.

Remark 2: The proof of Theorem 2 is straightforward and omitted here. Indeed, the proposed MSLR also encompasses the landmark work of multitask feature learning [5], which focuses on feature learning, as a special case with $\lambda = 0$. Compared with the OKL algorithms [30], [31], the MSLR is more generalized with a more relaxed sparsity constraint on U rather than on the regression matrix W, which allows to capture more complex, e.g., positive and negative, intertarget correlations rather than only similarity between output components in [30]. Moreover, we do not assume that all tasks are correlated and allow the existence of outlier tasks [52], which further increases the generality.

D. Solutions

The objective function (11) is not jointly convex with respect to $\boldsymbol{\alpha}$ and U. The nonsmoothness of the $\ell_{2,1}$-norm is generally regarded to be more difficult than the ℓ_1-norm minimization problems [50]. Therefore, it is challenging to solve for $\boldsymbol{\alpha}$ and U simultaneously. Fortunately, the objective is convex with each of $\boldsymbol{\alpha}$ and U when the other is given and fixed. We derive a new, fast alternating optimization algorithm to efficiently solve the objective function denoted by $F(\boldsymbol{\alpha}, U)$.

1) Fixing U to Solve for $\boldsymbol{\alpha}$: We calculate the gradients of the objective function with respect to $\boldsymbol{\alpha}$, and set to be $\mathbf{0}$

$$\frac{\partial F}{\partial \boldsymbol{\alpha}} = -\frac{1}{N}U^\top (Y - U\boldsymbol{\alpha}K)K + \lambda\boldsymbol{\alpha}K = \mathbf{0}. \quad (13)$$

We obtain

$$U^\top U\boldsymbol{\alpha}K + \lambda N\boldsymbol{\alpha} = U^\top Y. \quad (14)$$

Multiple K^{-1} to both sides on the right. We have

$$U^\top U\boldsymbol{\alpha} + \lambda\boldsymbol{\alpha}NK^{-1} = U^\top YK^{-1}, \quad (15)$$

which is a standard Sylvester equation

$$\mathcal{A}\Theta + \Theta\mathcal{B} = \mathcal{C} \quad (16)$$

where Θ is the unknown corresponding to $\boldsymbol{\alpha}$ in (15), $\mathcal{A} = U^\top U$, $\mathcal{B} = \lambda NK^{-1}$ and $\mathcal{C} = U^\top YK^{-1}$. Therefore, (15) has a closed-form solution and can be calculated efficiently for large scale problems by SLICOT[1] [30].

[1] www.slicot.org

Algorithm 1 Iterative Solution of U

Input: Data matrices X associated with corresponding outputs Y, regularization parameters λ and β.
Output: The structure matrix U.
1: Randomly initialize $U \in \mathbb{R}^{Q\times Q}$ and set $t = 1$;
2: **repeat**
3: Calculate the diagonal matrix $D^{(t)}$ using (19):
4: Calculate $U^{(t+1)}$ using (21):

$$U^{(t+1)} = YK\boldsymbol{\alpha}^\top(\boldsymbol{\alpha}KK\boldsymbol{\alpha}^\top + \beta ND^{(t)})^{-1}$$

5: $t \leftarrow t + 1$
6: **until** Convergence.
7: **return** U.

Algorithm 2 MSLR

Input: Data matrices X associated with corresponding outputs Y, regularization parameters λ and β.
Output: The regression coefficient matrix $\boldsymbol{\alpha}$ and the structure matrix U.
1: Randomly initialize $U \in \mathbb{R}^{Q\times Q}$ and set $i = 1$;
2: **repeat**
3: Calculate the matrix $\boldsymbol{\alpha}^{(i+1)}$ by solving the Sylvester equation in (15);
4: Calculate the $U^{(i+1)}$ using **Algorithm 1**;
5: $i \leftarrow i + 1$;
6: **until** Convergence.
7: **return** $\boldsymbol{\alpha}$ and U.

2) Fixing $\boldsymbol{\alpha}$ to Solve for U: We propose a very efficient iterative optimization algorithm to solve U. Taking the derivative of F with respect to U, we have

$$\frac{\partial F}{\partial U} = -2\frac{1}{N}(Y - U\boldsymbol{\alpha}K)(\boldsymbol{\alpha}K)^\top + 2\beta UD, \tag{17}$$

where D is a diagonal matrix with

$$(D)_{ii} = \frac{1}{2\|\mathbf{u}_i\|_2} \tag{18}$$

in which $\mathbf{u}_i$ is the ith row of $U^\top$. To avoid division by zero, we further regularize $(D)_{ii}$ by

$$(D)_{ii} = \frac{1}{2\sqrt{\mathbf{u}_i\mathbf{u}_i^\top + \zeta}} \tag{19}$$

where $\zeta > 0$ is a small constant. It is easy to check (19) approximates (18) when $\zeta \to 0$. Although the $\ell_{2,1}$-norm is nonsmooth, $F(\boldsymbol{\alpha}, U)$ is not differentiable only when a output is exactly predicted with a zero residual, which, however, is unlikely in real applications [9].

By setting the derivative to be $\mathbf{0}$, we obtain

$$\beta NUD + U\boldsymbol{\alpha}KK\boldsymbol{\alpha}^\top = YK\boldsymbol{\alpha}^\top. \tag{20}$$

U can be iteratively solved by

$$U = YK\boldsymbol{\alpha}^\top(\boldsymbol{\alpha}KK\boldsymbol{\alpha}^\top + \beta ND)^{-1}. \tag{21}$$

In each iteration, U is calculated with the current D and then D is updated based on the newly calculated U. The algorithm is summarized in Algorithm 1. The fast convergence of Algorithm 1 [50] guarantees the efficiency of the alternating optimization, which is summarized in Algorithm 2.

E. Convergence Analysis

The newly derived alternating optimization algorithm is computationally efficient with convergence which is guaranteed by Theorem 3. The rigorous proof has also been provided for theoretical completeness.

Theorem 3: $F(\boldsymbol{\alpha}, U)$ in (11) is bounded from the following and monotonically decreases with each optimization step for $\boldsymbol{\alpha}$ and U, and therefore, it converges.

Proof: Since $F(\boldsymbol{\alpha}, U)$ is the summation of norms, we have $F(\boldsymbol{\alpha}, U) \geq 0$ for any $\boldsymbol{\alpha}$ and U. Then, $F(\boldsymbol{\alpha}, U)$ is bounded from the following. Denote $\boldsymbol{\alpha}^{(i)}$ and $U^{(i)}$ as the $\boldsymbol{\alpha}$ and U in the ith iteration, respectively. For the ith step, $\boldsymbol{\alpha}^{(i)}$ is computed by $\boldsymbol{\alpha}^{(i)} \leftarrow \arg\min_{\boldsymbol{\alpha}} F(\boldsymbol{\alpha}, U^{(i-1)})$. It has been proved in [50] that $F(\boldsymbol{\alpha}^{(i)}, U^{(i-1)}) \geq F(\boldsymbol{\alpha}^{(i)}, U^{(i)})$. We therefore have the following inequality:

$$\begin{aligned}\cdots \geq F(\boldsymbol{\alpha}^{(i-1)}, U^{(i-1)}) \geq F(\boldsymbol{\alpha}^{(i)}, U^{(i-1)}) \\ \geq F(\boldsymbol{\alpha}^{(i)}, U^{(i)}) \geq \cdots.\end{aligned}$$

Therefore, $F(\boldsymbol{\alpha}^{(i)}, U^{(i)})$ monotonically decreases as $i \to +\infty$, which indicates that the objective function $F(\boldsymbol{\alpha}, U)$ converges according to the monotone convergence theorem. □

The convergence of the alternating optimization algorithm is theoretically important and ensures the efficiency and effectiveness of the MSLR for multitarget regression. It is also practically meaningful for its wide use in real applications.

F. Complexity Analysis

The complexity of the proposed alternating optimization algorithm stems from solving for α and U in Algorithms 1 and 2, respectively. Similar to existing kernel methods, e.g., kernel ridge regression (KRR), the computation of inversion Gram matrix is of complexity $\mathcal{O}(N^3)$, where N is the number of training samples. Assume the iteration steps of Algorithms 1 and 2 are t_1 and t_2, respectively, and then, the total complexity of the alternating optimization algorithm is $\mathcal{O}(t_1t_2N^3) + \mathcal{O}(t_2N^3)$ which is approximately $\mathcal{O}(N^3)$ due to the fact that $t_1(\approx 5) \ll N$ and $t_2(\approx 15) \ll N$. Therefore, the complexity of the proposed MSLR is approximately the same as regular kernel methods, e.g., KRR, which could be sped up by sparsification [48]. The efficiency of computing the Gram matrix is ensured by recent advances on kernel approximation methods, e.g., random Fourier features [53], which allow to scale well with very large data sets.

IV. Experiments and Results

We demonstrate the great effectiveness of the MSLR for diverse multitarget regression tasks on both synthetic data and real-world data sets, and provide experimental analysis of convergence.

To directly compared with previous methods, we employ the commonly used relative root mean squared error (RRMSE) as

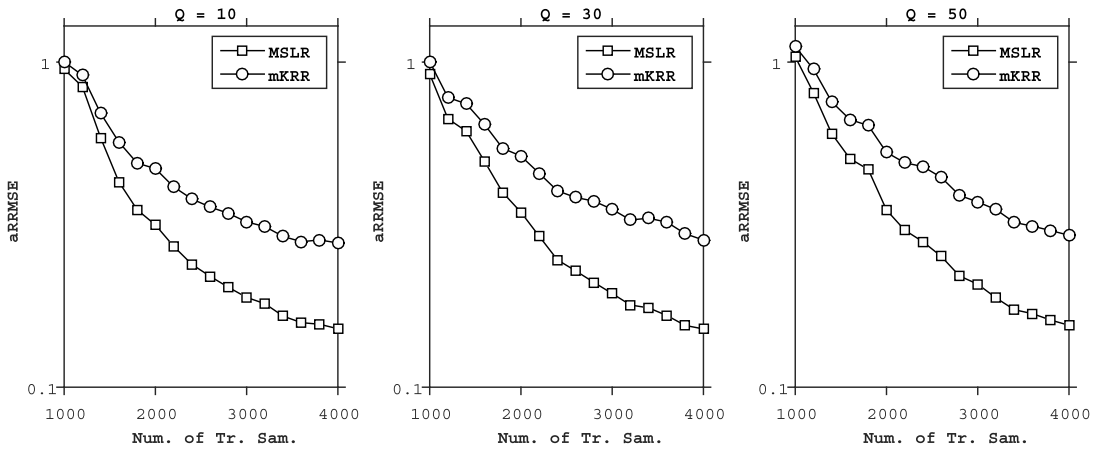


Fig. 3. Performance of the proposed MSLR versus mKRR on synthetic data with different numbers of training samples. Q is the number of outputs.

the measurement of performance. The RRMSE is defined as

$$\text{RRMSE} = \sqrt{\frac{\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in D_{\text{test}}}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^2}{\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in D_{\text{test}}}(\hat{Y} - \mathbf{y}_i)^2}},$$

where $(\mathbf{x}, \mathbf{y}_i)$ is the ith sample $\mathbf{x}$ with ground truth output $\mathbf{y}_i$, $\hat{\mathbf{y}}_i$ is the prediction of $\mathbf{y}_i$, and $\hat{Y}$ is the average of the outputs over the training set D_{train}. We take the average RRMSE (aRRMSE) across all the output variables within the test set D_{test} as a single measurement. The aRRMSE is the relative errors to those by predicting all outputs as the means of training samples. A lower aRRMSE indicates better performance. The parameters λ and β are obtained by cross validation on the training set by fixing one to tune the other. We use the radial basis function (RBF) kernel for nonlinear regression, and the bandwidth is set to be the mean of pairwise distances of all training samples, which automatically adapts to different data sets [49], [54]. The code will be released with the authors' Web page.

A. Experiments on Synthetic Data

To show the ability of the proposed MSLR to jointly model intertarget correlations and nonlinear input–output relationships, we provide the evaluation on synthetic data. We conduct extensive experiments on synthetic data for multitarget regression tasks of different numbers of outputs. To the show the oracle performance, we have also experimented with data generated by the true model. For contrast, we compare the proposed MSLR with the baseline multitarget KRR (mKRR), which does not take into account the intertarget correlations on synthetic data.

1) Data Sets: We adopt the simulation methods in prior work [8], [16], [28], [55] to generate nonlinear multitarget regression by $\mathbf{y}_i = W\phi(\mathbf{x}_i) + \boldsymbol{\epsilon}$, where $\mathbf{y}_i \in \mathbb{R}^Q$, $\mathbf{x}_i \in \mathbb{R}^d$ is drawn from multivariate Gaussian distributions, $\phi(\mathbf{x}_i) = (\mathbf{x}_i^2, \mathbf{x}_i, 1) \in \mathbb{R}^{2d+1}$ is the feature map of $\mathbf{x}_i$, and $\boldsymbol{\epsilon}$ is the added Gaussian noise. We introduce the correlations between outputs by $W \in \mathbb{R}^{Q\times(2d+1)}$ generated by $W = U^\top U W_0$, where $U \in \mathbb{R}^{q\times Q}$ and $W_0 \in \mathbb{R}^{Q\times(2d+1)}$ are randomly created and $q < Q$. We set $d = 60$ and experiment with different number of outputs, i.e., $Q = 10, 30$, and 50. We generate 100 samples for test and up to 4000 samples for training.

TABLE I

PERFORMANCE (aRRMSE) OF THE PROPOSED MSLR VERSUS THE mKRR ON THE SYNTHETIC DATA GENERATED BY THE TRUE MODEL. Q IS THE NUMBER OF OUTPUTS

Method \ Target (Q)	10	20	30	40	50
MSLR	0.0052	0.0017	0.0013	0.0012	0.0007
mKRR	0.0376	0.0376	0.0384	0.0348	0.0373

To further show the advantages of the proposed MSLR, we have also experimented on the synthetic data generated by the true model. Specifically, we generate input data $X \in \mathbb{R}^{d\times N}$, parameters $\boldsymbol{\alpha} \in \mathbb{R}^{Q\times N}$ and a sparse $U \in \mathbb{R}^{Q\times Q}$. The outputs $Y \in \mathbb{R}^{Q\times N}$ are computed by $Y = U\boldsymbol{\alpha}K$, which K is the Gram matrix on data X and computed with the RBF kernel. We experiment on a set of 4000 and 100 samples for training and test, respectively, with a large range of numbers ($Q = 10$, 20, 30, 40, and 50) of outputs.

2) Performance: Fig. 3 shows the performance comparison between the proposed MSLR and the mKRR with different numbers of training samples. The mKRR decouples multitarget regression into several independent single output regression tasks without considering intertarget correlations. The proposed MSLR consistently outperforms mKRR in terms of the aRRMSE with different numbers of training samples. Note that both the mKRR and the proposed MSLR perform better with the increase of the number of training samples, while the proposed MSLR can produce much larger performance improvement over the mKRR, because correlations are better encoded with more training samples, which shows the effectiveness of the MSLR in modeling intertarget correlations. The proposed MSLR shows its great capability of jointly modeling intertarget correlations and nonlinear input–output relationships on synthetic data. The linear regression, e.g., mKRR with a linear kernel, cannot handle nonlinear input–output relationships and produces very poor results (aRRMSE $>$ 1), and therefore, the result is not plotted in Fig. 3.

The comparison of the proposed MSLR with the mKRR on the synthetic data generated by the true model is reported in Table I. Our MSLR can produce highly accurate prediction with very low aRRMSE especially with number of outputs of

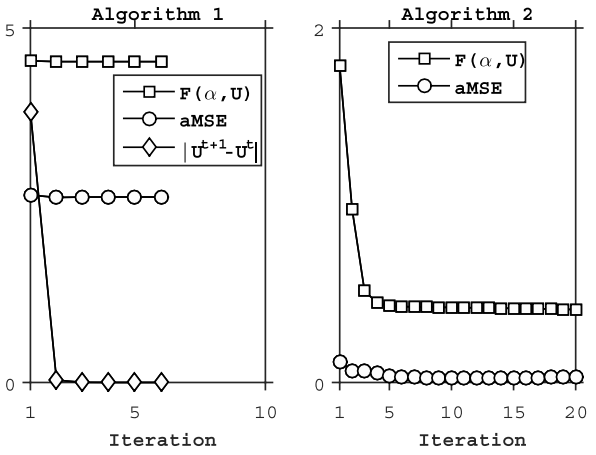


Fig. 4. Convergence illustration of the proposed algorithms on synthetic data. $F(\boldsymbol{\alpha}, U)$ is the objective function. $|U^{t+1} - U^t|$ is the norm of the difference between U^{t+1} and U^t. aMSE denotes the average mean square error of multiple outputs

$Q = 50$, which shows the benefits of the introduced latent space. The mKRR can achieve good performance on this data set, which is still much lower than that of our MSLR. The performance on this data set indicates the great effectiveness of our MSLR in modeling intertarget correlations to improve the performance of multitarget regression.

3) Convergence: Fig. 4 shows the convergence of the propose Algorithms 1 and 2 on synthetic data. We omit experiments with $Q = 10$ to avoid redundance. The algorithms converge very quickly with a few iterations, especially for Algorithm 1, which can converge within $2 \sim 3$ steps. The quick convergence of both Algorithm 1 and 2 shows the effectiveness and efficiency of the proposed algorithms in solving the objective function and guarantees the efficiency of the proposed MSLR for multitarget regression.

B. Experiments on Real-World Data Sets

The effectiveness and the generality of the proposed MSLR have been experimentally validated by the high performance on six challenging and diverse real-world data sets, which are recently released and publicly available [37].

1) Data Sets: The six data sets are greatly diverse covering a broad range of multitarget prediction tasks from price prediction to river flow estimation, which are challenging for multitarget regression models. The statistics of the six data sets are summarized in Table II. We have also plotted the correlation coefficients between multiple outputs on the six data sets in Fig. 5. We can observe that on all the six data sets multiple outputs demonstrates strong correlations which can be explored for improved, more accurate multitarget regression. Note that the correlations demonstrate diverse patterns according to different data sets, which poses great challenges and requires learning from data.

a) ATP1d & ATP7d: The Airline Ticket Price data set concerns the prediction of airline ticket prices. The inputs, which are a feature set of 411 variables, contain the values that may be useful for prediction of the airline ticket prices for a specific departure date. The output variables are prices of the next day (ATP1d) or minimum price observed over the next seven days (ATP7d) for six output flight preferences.

TABLE II

STATISTICS OF THE SIX DATA SETS. d AND Q DENOTE DIMENSIONS OF THE INPUT AND OUTPUT, RESPECTIVELY

Dataset	Samples (train/test)	Input (d)	Target (Q)
ATP1d	337	411	6
ATP7d	296	411	6
RF1	4108/5017	64	8
RF2	4108/5017	576	8
SCM1d	8145/1658	280	16
SCM20d	7463/1503	61	16

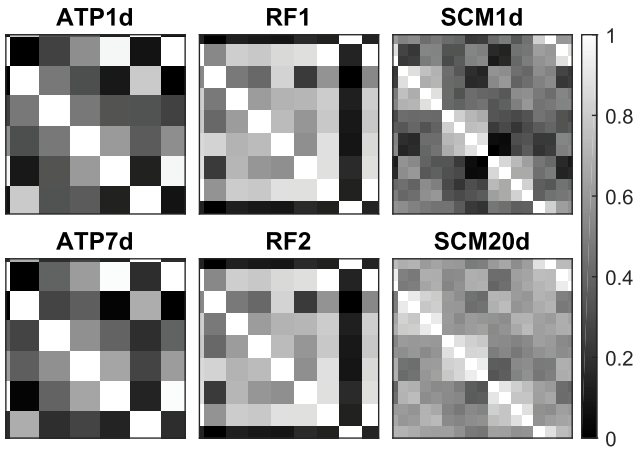


Fig. 5. Illustration of correlation between outputs on the six different data sets. Higher intensity indicates higher correlation in the right bar.

b) RF1 & RF2: The River Flow data sets concern the prediction of river network flows for 48 h in the future at specific locations. The data set contains data from hourly flow observations for eight sites in the Mississippi River network in the USA and was obtained from the U.S. National Weather Service. In RF1, each site contributes eight attribute variables to facilitate prediction. There are a total of 64 variables plus eight output variables. The RF2 data set extends RF1 by adding precipitation forecast information for each of the eight sites.

c) SCM1d & SCM20d: The Supply Chain Management data sets are derived from the Trading Agent Competition in Supply Chain Management (TAC SCM) tournament from 2010. The input variables are observed prices for a specific tournament day. The 16 regression outputs, each output corresponds to the next day mean price (SCM1d) or mean price for 20 days in the future (SCM20d).

2) Implementation Settings: We follow the validation settings in [34] for each data set to benchmark with other algorithms. We compare the state-of-the-art algorithms in [8], [32], [34], and [37], including single task learning (STL), multiobject random forests (MORF), the corrected multitask stacking, ERC, random linear target combinations (RLC) [32]. We have also implemented representative multitarget regression models including the mSVR [49], mKRR, adaptive k-cluster random forests (AKRF) [6], OKL [31], MTRL [16], and MROTS [8]. For the mSVR and mKRR, the bandwidth of the Gaussian kernel is set to the mean of all pairwise distances of training

TABLE III
COMPARISON OF THE PROPOSED MSLR WITH THE STATE-OF-THE-ART ALGORITHMS ON SIX DATA SETS IN TERMS OF aRRMSE (%)

Dataset \ Method	STL	MTSC	ERCC	RLC	MORF	mSVR	AKRF	MROTS	OKL	MTFL	MTRL	mKRR	**MSLR**
ATP1d	37.35	37.17	37.24	38.40	42.22	38.10	41.22	40.37	36.44	41.54	37.32	37.98	**35.17**
ATP7d	52.48	50.74	51.24	46.14	55.08	47.75	53.08	54.90	47.53	55.33	50.14	48.56	**45.77**
RF1	69.63	69.82	69.89	72.65	85.13	75.43	85.56	81.40	74.26	82.34	77.21	79.20	**66.26**
RF2	69.64	69.86	69.82	70.36	91.89	83.56	90.32	81.76	77.54	82.86	78.45	84.79	**67.14**
SCM1d	47.75	47.01	46.63	45.72	56.63	54.28	55.84	47.59	48.34	47.61	46.79	48.64	**43.34**
SCM20d	77.68	78.54	75.97	74.67	77.75	76.28	77.65	76.40	76.98	77.56	77.11	77.87	**74.02**

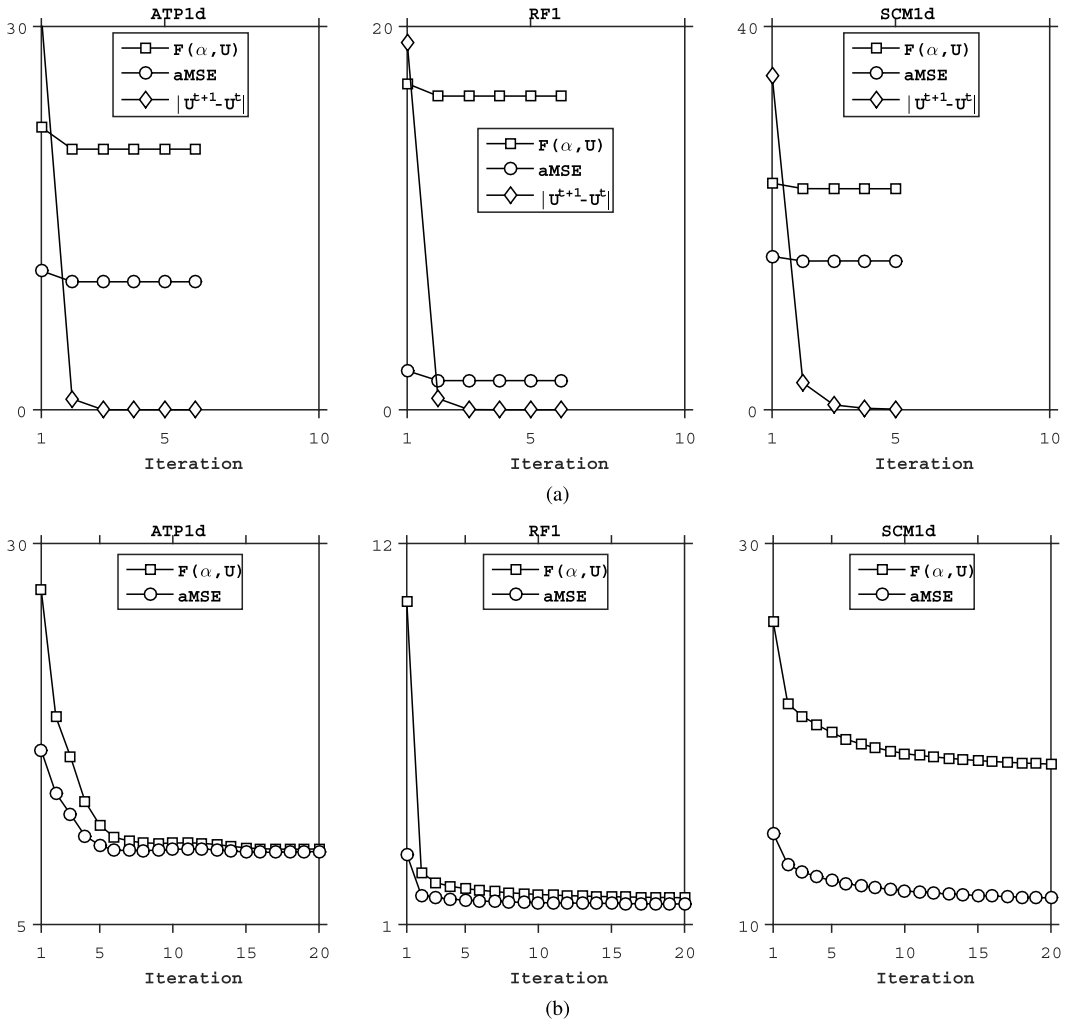


Fig. 6. Convergence illustration of the proposed algorithms. $F(\boldsymbol{\alpha}, U)$ is the objective function. $|U^{t+1} - U^t|$ is the norm of the difference between U^{t+1} and U^t. aMSE denotes the average mean square error of multiple outputs. (a) Convergence of Algorithm 1 to update U. (b) Convergence of Algorithm 2 for alternating optimization.

samples, which generally produces the best performance [54]; the parameters C and ϵ in the mSVR are fixed to be 1000 and 0.0001, respectively, by following the original work. For the AKRF, following the original work, we train 20 trees to build the regression forests, set the splitting number to be 2, use the minimum leaf size of 5, which means that a node is not split any more if the number of training instances associated with the node is less than or equal to 5; the splitting is conducted by a linear support vector machine [56] in which the parameters are set by default as in [6]; the sample rate is set to be 1 for training each tree. For the OKL, we fix the parameter λ to be 10^2, since it produces consistently optimal performance. For the MROTS, according to the experimental results in the original work, we set the iteration to be 50; λ is fixed to be 0.0001; λ_1 and λ_2 are selected by cross validation in the grid of $10^{[-3:1:3]}$. Similarly, for the MTRL, the parameter λ_1 and λ_2 are selected by cross validation in the grid of $10^{[-3:1:3]}$.

3) Performance: The proposed MSLR algorithm has achieved consistently high prediction performance on all six data sets and substantially outperforms the state-of-the-art algorithms developed recently. The comparison results in terms of aRRSM are summarized in Table III.

1) The proposed MSLR substantially outperforms the best results from the state-of-the-art algorithms on all the six data sets by large margins up to 4.84%. The high accuracy on the six diverse data sets validates the effectiveness of the MSLR for a broad range of multitarget prediction tasks. The consistently high performance of MSLR on all the six data sets of huge diversity shows the generality of MSLR in extracting intertarget correlations automatically from data, which indicates its wide use in broad applications.
2) The large improvement of the proposed MSLR over the STL/mKRR methods with significant margins on the all the six data sets by up to 12% and 20%, respectively, showing its effectiveness in jointly modeling intertarget correlations and input–output relationships. The STL [37] and mKRR are regarded as the baseline methods that predict multiple outputs independently without exploring the correlation among multiple outputs. The consistently better performance of MSLR than mKRR indicates the effectiveness of MSLR in disentangling the nonlinear relationship between inputs and outputs while capturing intrinsic intertarget correlations by automatically inferring from data for diverse applications.
3) The advantage of the proposed $\ell_{2,1}$-norm-based sparse learning of the structure matrix on modeling the intertarget correlation has been demonstrated by the large improvement up to 26% over these methods, which model intertarget correlations including MTSC, ERCC, RLC, mSVR, AKRF, MROTS, MTFL, MTRL, and OKL. The much better performance of the MSLR than these methods also shows its generality of modeling the correlations for diverse tasks. The OKL deploys a similarity kernel to model the correlation of multiple output, which lacks the effectiveness to fully capture complex intertarget correlations, including both positive and negative correlations. Although kernel extension to achieve nonlinear multitarget regression is provided, the MTRL is not able to effectively capture the intertarget correlations due to the strong assumptions. The large performance improvement over the MTFL, MTRL, and OKL indicates the advantages of the MSLR in jointly handling input–output relationships and intertarget correlations.

4) Convergence: Both Algorithms 1 and 2 converge very fast within a few iterations, which enables efficient multitarget regression. We show convergence results on three representative data sets: ATP1d, RF1, and SCM1d. As shown in Fig. 6(a), Algorithm 1 converges very quickly within a few iteration steps on all the data sets. We check the norm of the difference between $U^{(t+1)}$ and $U^{(t)}$, namely, $||U^{(t+1)} - U^{(t)}|| < 0.0001$ as the criterion for convergence, which reaches 0 very fast very few (2∼3) iterations. The quick convergence of Algorithm 1 ensures to efficiently update U in solving the alternating optimization of Algorithm 2. As shown in Fig. 6(b), Algorithm 2 converges quickly within 20 iteration steps, which guarantees its effectiveness and efficiency for multitarget regression. We check $F^{(t)}(\alpha, U) - F^{(t+1)}(\alpha, U) < 0.001$ as the stopping criterion for convergence.

The high performance on both synthetic data and real data sets shows the effectiveness and generality of the proposed MSLR in jointly modeling intertarget correlations and handling nonlinear input–output relationships. The great advantages over previous multitarget regression models demonstrate the strength of the MSLR as a latent variable model for multitarget regression. The fast convergence of alternating optimization guarantees the computational efficiency of the proposed MSLR in practical applications.

V. Conclusion

We have presented a novel multitarget regression model called MSLR, which enables simultaneously modeling intrinsic intertarget correlations and highly complex nonlinear input–output relationships in one single framework. By incorporating a latent space, the MSLR can explicitly encode the intertarget correlations in a structure matrix, which avoids the reliance on specific assumptions required in previous work. Thanks to the incorporation of the latent space associated with the structure matrix, the MSLR naturally admits a linear representer theorem, which enables kernel extension for nonlinear multitarget regression. The MSLR accomplishes a novel latent variable model of multitarget regression, which indicates the power of the multilayer learning architecture for multitarget regression. Experimental results on both synthetic data and six diverse real-world data sets have shown that the MSLR achieves high performance and largely outperforms the state-of-the-art representative algorithms, which validates its great effectiveness for diverse multitarget regression tasks.

References

[1] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[2] K. Slavakis, P. Bouboulis, and S. Theodoridis, "Adaptive multiregression in reproducing kernel Hilbert spaces: The multiaccess MIMO channel case," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 2, pp. 260–276, Feb. 2012.
[3] G. Skolidis and G. Sanguinetti, "Semisupervised multitask learning with Gaussian processes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 2101–2112, Dec. 2013.
[4] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, Nov. 2005.
[5] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.
[6] K. Hara and R. Chellappa, "Growing regression forests by classification: Applications to object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 552–567.
[7] A. J. Rothman, E. Levina, and J. Zhu, "Sparse multivariate regression with covariance estimation," *J. Comput. Graph. Statist.*, vol. 19, no. 4, pp. 947–962, 2010.
[8] P. Rai, A. Kumar, and H. Daume, "Simultaneously leveraging output and task structures for multiple-output regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3185–3193.
[9] H. Liu, L. Wang, and T. Zhao, "Multivariate regression with calibration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 127–135.
[10] Y. Zhang and D.-Y. Yeung, "Learning high-order task relationships in multi-task learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1917–1923.
[11] F. Dinuzzo and B. Schölkopf, "The representer theorem for Hilbert spaces: A necessary and sufficient condition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 189–196.
[12] K. Yu, V. Tresp, and A. Schwaighofer, "Learning Gaussian processes from multiple tasks," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 1012–1019.

[13] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, "Learning a meta-level prior for feature relevance from multiple related tasks," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 489–496.

[14] A. Kumar and H. Daume, "Learning task grouping and overlap in multi-task learning," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1383–1390.

[15] A. Agarwal, S. Gerber, and H. Daumé, III, "Learning multiple tasks using manifold regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 46–54.

[16] Y. Zhang and D.-Y. Yeung, "A regularization approach to learning task relationships in multitask learning," *ACM Trans. Knowl. Discovery Data*, vol. 8, no. 3, p. 12, Jun. 2014.

[17] G. S. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *Ann. Math. Statist.*, vol. 41, no. 2, pp. 495–502, Apr. 1970.

[18] C. Ding, D. Zhou, X. He, and H. Zha, "R_1-PCA: Rotational invariant L_1-norm principal component analysis for robust subspace factorization," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 281–288.

[19] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 41–48.

[20] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 339–348.

[21] P. Gong, J. Ye, and C. Zhang, "Multi-stage multi-task feature learning," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2979–3010, 2013.

[22] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 4, p. 22, Feb. 2012.

[23] P. Jawanpuria and J. S. Nath, "A convex feature learning formulation for latent task structure discovery," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1–8.

[24] Q. Zhou and Q. Zhao, "Flexible clustered multi-task learning by learning representative tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 266–278, Feb. 2016.

[25] F. Cai and V. Cherkassky, "Generalized SMO algorithm for SVM-based multitask learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 997–1003, Jun. 2012.

[26] C. Ciliberto, Y. Mroueh, T. Poggio, and L. Rosasco, "Convex learning of multiple tasks and their structure," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1548–1557.

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

[28] K.-A. Sohn and S. Kim, "Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 1081–1089.

[29] M. Álvarez, L. Rosasco, and N. Lawrence, *Kernels for Vector-Valued Functions: A Review* (Foundations and Trends in Machine Learning). London, U.K.: Now publishers, 2012.

[30] F. Dinuzzo, C. S. Ong, G. Pillonetto, and P. V. Gehler, "Learning output kernels with block coordinate descent," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 49–56.

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

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

[33] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, "Ensembles of multi-objective decision trees," in *Proc. Eur. Conf. Mach. Learn.*, 2007, pp. 624–631.

[34] G. Tsoumakas, E. Spyromitros-Xioufis, A. Vrekou, and I. Vlahavas, "Multi-target regression via random linear target combinations," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 225–240.

[35] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Advances in Knowledge Discovery and Data Mining*. 2004, pp. 22–30.

[36] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *J. Mach. Learn.*, vol. 85, no. 3, pp. 333–359, Dec. 2011.

[37] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas. (2014). "Multi-target regression via input space expansion: Treating targets as inputs." [Online]. Available: https://arxiv.org/abs/1211.6581

[38] A. Bargi, M. Piccardi, and Z. Ghahramani, "A non-parametric conditional factor regression model for multi-dimensional input and response," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2014, pp. 77–85.

[39] L. Jacob, J.-P. Vert, and F. R. Bach, "Clustered multi-task learning: A convex formulation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 745–752.

[40] C. Ciliberto, L. Rosasco, and S. Villa, "Learning multiple visual tasks while discovering their structure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2015, pp. 131–139.

[41] X. Zhen, Z. Wang, M. Yu, and S. Li, "Supervised descriptor learning for multi-output regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1211–1218.

[42] J. Charles, T. Pfister, M. Everingham, and A. Zisserman, "Automatic and efficient human pose estimation for sign language videos," *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 70–90, 2014.

[43] X. Zhen, Y. Yin, M. Bhaduri, I. B. Nachum, D. Laidley, and S. Li, "Multi-task shape regression for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent. (MICCAI)*, 2016, pp. 210–218.

[44] X. Zhen, H. Zhang, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Direct and simultaneous estimation of cardiac four chamber volumes by multioutput sparse regression," *Med. Image Anal.*, vol. 36, pp. 184–196, Feb. 2017, doi: 10.1016/j.media.2016.11.008.

[45] J. Gillberg *et al.* (2014). "Multiple output regression with latent noise." [Online]. Available: https://arxiv.org/abs/1410.7365

[46] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[47] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sep. 2001.

[48] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16. 2004, pp. 329–336.

[49] 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," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2298–2307, Aug. 2004.

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

[51] 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 *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1466–1474.

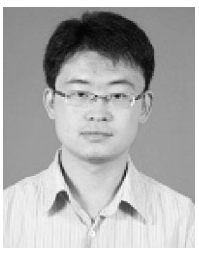
[52] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 895–903.

[53] B. Sriperumbudur and Z. Szabó, "Optimal rates for random Fourier features," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1144–1152.

[54] X. Zhen, M. Yu, A. Islam, M. Bhaduri, I. Chan, and S. Li, "Descriptor learning via supervised manifold regularization for multioutput regression," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

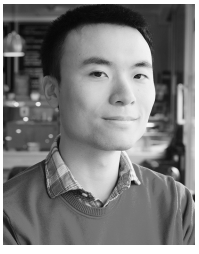
[55] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2010, pp. 733–742.

[56] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.



Xiantong Zhen received the B.S. and M.E. degrees from Lanzhou University, Lanzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree from the Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, U.K., in 2013.

He is currently a Post-Doctoral Fellow with the University of Western Ontario, London, ON, Canada. His current research interests include machine learning, computer vision, and medical image analysis.



Mengyang Yu (S'14) received the B.S. and M.S. degrees from the School of Mathematical Sciences, Peking University, Beijing, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, U.K.

His current research interests include computer vision, machine learning, and data mining.

Feng Zheng received the B.S. and M.S. degrees in applied mathematics from Hubei University, Wuhan, China, in 2006 and 2009, respectively, and the Ph.D. degree from the Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, U.K.

From 2009 to 2012, he was with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China, as an Assistant Researcher, from 2009 to 2011, and an Assistant Research Professor, from 2011 to 2012. He is currently a Research Fellow with The University of Texas at Arlington, Arlington, TX, USA. His current research interests include computer vision, machine learning, and human–computer interaction.

Ilanit Ben Nachum, photograph and biography not available at the time of publication.

Mousumi Bhaduri, photograph and biography not available at the time of publication.

David Laidley, photograph and biography not available at the time of publication.

Shuo Li received the Ph.D. degree in computer science from Concordia University, Montrál, QC, Canada, in 2006.

He was a Research Scientist and a Project Manager of General Electric (GE) Healthcare, London, ON, Canada, for nine years, where he is currently an Associate Professor with the Department of Medical Imaging and Medical Biophysics, and also a Scientist with the Lawson Health Research Institute, London. He is the Founder and has been the Director of the Digital Imaging Group of London, London, since 2006, which is a highly dynamic and interdisciplinary group. He has authored or co-authored over 100 publications and edited five Springer books. His current research interests include the development of intelligent analytic tools to facilitate physicians and hospital administrative to handle the big medical data, centered with medical images.

Dr. Li was a recipient of several GE internal awards. His Ph.D. thesis received the Doctoral Prize giving to the most deserving graduating student in the Faculty of Engineering and Computer Science. He serves as a Guest Editor and an Associate Editor of several prestigious journals. He served as a Program Committee Member in top conferences.