Intra- and Inter-sparse Multiple Output Regression with Application on Environmental Microbial Community Study

Jie Yang†, Henry C.M. Leung†, S.M. Yiu†, Yunpeng Cai‡, Francis Y.L. Chin†*
† Department of Computer Science
The University of Hong Kong, Hong Kong
‡ Research Center for Biomedical Information Technology
Shenzhen Institutes of Advance Technology & Key Lab for Health Informatics
Chinese Academy of Sciences, China

Abstract—**Feature selection is important for many biological studies, especially when the number of available samples is limited (in order of hundreds) while the number of input features is large (in order of millions), such as eQTL (expression quantitative trait loci) mapping, GWAS (genome wide association study) and environmental microbial community study. We study the problem of multiple output regression which leverages the underlying common relationship shared by multiple output features and propose an efficient and accurate approach for feature selection. Our approach considers both intra- and inter- group sparsities. The intergroup sparsity assumes that only small set of input features are related to the output features. The intragroup sparsity assumes that each input features may relate to multiple output features which should have different kinds of sparsity. Most existing methods do not model the intragroup sparsity well by either assuming uniform regularization on each group, i.e. each input feature relates to similar number of output features, or requiring prior knowledge of the relationship of input and output features. By modelling the regression coefficients as a mixture distributions of Laplacian and Gaussian, we can shrink group regression coefficients to be small adaptively and learn the intergroup, intragroup sparsity and shrinkage estimation patterns. Empirical studies on the synthetic and real environmental microbial community datasets show that our model has better predictions on test dataset than existing methods such as Lasso, Elastic Net, dirty model and rMTFL (robust multi-task feature learning). Moreover, by using least angle regression or coordinate descent and projected gradient descent techniques for optimization, we can obtain the optimal regression efficiently.**

I. INTRODUCTION

Many biological studies focus on finding a subset of features from a large set of features related to particular set of observations, for example, eQTL (expression quantitative trait loci) mapping aims at finding loci of the genome related to gene expression levels [10]; GWAS [1] (genome wide association study) examines many common genetic variants in different individuals to see if any variant is associated with a trait; and environmental microbial community study wants to discover sets of microbes sensitive to different environmental features. The most popular method to model this kind of problem is to fit a linear regression model from input features X (loci of genome, genetic variants, microbes communities) to output features Y (gene expression levels, different traits, environmental factors). By learning the regression coefficients B from $Y = XB + \varepsilon$ where ε is the noise matrix, the non-zero coefficients are interpreted as potential interactions between input and output features. There are two main reasons why we consider linear regression:

- Linear model is less likely to overfit than non-linear models;
- Linear model is easier to be interpreted than other models.

However, in biological studies, the number of available samples is usually much less than the number of input features. For example, there can be thousands of loci, hundred thousands of genetic variants and millions of microbes for the eQTL mapping, GWAS and environmental microbes community study respectively but the number of samples are usually less than one thousand. Therefore, the existing learning model is likely to encounter overfitting problem when applying on biological data because the covariance matrix $X^T X$ might be singular. Luckily, the number of input features relates to a particular out feature is usually small in biological studies. For example, only a few loci may relate to the expression of a gene; about hundreds of genetic variants may relate to a trait; only about hundreds of microbes sensitive to a particular environment. To overcome the overfitting problem, regularizations are employed while learning from data with an assumption that only a small number of input features are related to output features.

When there is only one output feature, one regularization approach is to restrict the number of non-zero regression coefficients by including $||B||_0$ (the L_0 of the regression coefficients) as penalty. However, this linear regression problem with L_0 penalty is NP-hard [26]. Lasso [26] proposes another regularization approach to impose sparsity on the regression coefficients by approximating L_0 using L_1, i.e using $||B||_1$ as penalty. This approach can generate sparse solutions, equivalent to select a small number of input features (*feature selection*). However, when there are many highly dependent input features, Lasso is sensitive to the noise matrix and has bad performance. Ridge regression [18] employs $||B||_2$ (the L_2 of the regression coefficient) for regularization, which tends to reduce the value of regression coefficients (*shrinkage estimation*) to avoid overfitting [18] with the increase of the number of the non-zero coefficients. Elastic Net [30], which employs both L_1 and L_2 penalty for regularization has the advantages of both Lasso and Ridge, i.e. simultaneous feature selection and shrinkage estimation. All the above algorithms assume each input feature is independent. In real situation,

* to whom correspondence should be addressed. Email: chin@cs.hku.hk

some input features may be related and they should have similar coefficients. For example, loci and genetic variants located closely in the genome are more likely related to the expression of similar set of genes and similar traits respectively than loci and genetic variants located far away. Microbes usually live together are likely to be sensitive to similar environmental features. Group Lasso [29] and its extension [15] assign input features with similar values (in all samples) into groups and grouped input features should be selected or injected together, i.e. with all non-zero or zero coefficients on the grouped input features. However, they fail to consider those not-so-similar features.

When there are multiple output features, the above algorithms can be applied for each output feature independently. However, by considering all output features together, better regression could be obtained because

- some output features may be associated and should have similar regression coefficients;
- some input features may be highly related to a subset of output features.

For example, genes with similar expression profiles should be related to similar loci, microbes live in deep sea should be sensitive to temperature, too. Models may benefit from this shared common relationship and prevent overfitting [20], [22], [24]. [9], [20] use trees to model the relationship of input and output features. Although these algorithms work well when proper tree structures are available, the prior knowledge of the tree structures is difficult to be determined in practice and thus limits the applicability of these algorithms. [3], [4], [17], [19], [22] model the shared common relationship without prior knowledge by minimizing the number of non-zero entries in matrix B, i.e. the number of rows with non-zero entries (*intergroup sparsity*) and the number of non-zero entries in each row (*intragroup sparsity*). Although no prior knowledge of the data is required by these methods, the performances of these methods are not good as the intragroup sparsity is treated uniformly, i.e. each input feature is assumed to be related to similar number of output features. In real situation, each input feature may be related to different number of output features, e.g. genetic variants in transcription factor gene should relate to more traits than genetic variants in non-gene region, microbes appear in specific environment should be more sensitive to multiple environmental features than microbes appear globally. Thus different rows of matrix B may have different kinds of sparsity and should be considered separately. [5] studies the eQTL problem and considers adaptive penalties for different rows of matrix B. It requires the prior knowledge about genomic locations such as conservation scores and transcription factor binding sites.

In summary, for the regression problem with multiple output features, the existing algorithms either assume each input feature should be related to similar number of output features (uniform intragroup sparsity) or require detailed prior knowledge of the relationship of input features and association of output features. To solve this problem, our contributions are to have non-uniform intragroup sparsity for matrix B and to learn the intragroup sparsity from the input X and Y without any prior knowledge. Different intragroup regularization parameters are assigned for each row of matrix B with the assumption that the regression coefficients in different rows of matrix B should follow different Laplacian distributions and our model can learn the intragroup sparsity regularization parameters adaptively from the data, instead of cross validation which is impractical when there are many rows of matrix B. At the same time, in order to reduce the unstable feature selection caused by highly dependent input features, we consider adaptive shrinkage estimation for rows of matrix B and the shrinkage regularization parameters can be learned adaptively from the data by assuming different Gaussian distributions on different rows of matrix B. Also our penalty is modeled in a probability way by considering mixtures of Laplacian and Gaussian distributions on the regression coefficients. Our model can learn the intergroup sparsity simultaneously by only selecting a small sets of input features which are related to at least one output features. By applying our model to the simulated data of different sparsity levels, our model is able to learn sparsity, shrinkage patterns and recover true related input features with higher probability compared to existing methods such as Lasso, Elastic Net, dirty model [19] and rMTFL [17]. Also we apply our model to the global microbes communities to study the interaction patterns, our model has better prediction ability on microbes community data for both family cut and genus cut datasets.

The rest of paper are organized as follows. In Section II, we give a brief review of background and formulate our problem. Later we propose our method for adaptive sparsity learning and shrinkage estimation. In Section III, we present the empirical results on both simulated and real applications followed by the conclusion in Section IV.

II. Method

Background and Problem Formulation

Let X be an $N \times P$ matrix for N samples and P input features, where each element x_{ij} represents the value of the j^{th} input feature in the i^{th} sample. Let Y be an $N \times Q$ matrix representing N samples and Q output features where $y_{ik} \in R$ denotes the value of k^{th} output feature in the i^{th} sample. We assume that all samples are independent and the feature values are represented by identically distributed random variables. A univariate linear regression method can be employed to model the relationship between the input features X and each output feature y_k separately as:

$$y_k = XB_{*k} + \varepsilon_k, \forall k = 1,..,Q; \tag{1}$$

where B_{*k} is a length-P vector of regression coefficients for the k^{th} output feature. Let $B = \{B_1,...,B_Q\}$ and ε_k be the error term of length N with zero mean and constant variance following Gaussian distribution. We denote the j^{th} row and k^{th} column of matrix B as B_{j*} and B_{*k} respectively for convenience. We can obtain an optimal B by minimizing the residual sum of squares:

$$\hat{B} = argmin_B \sum_{k=1}^{Q} \frac{1}{2}(y_k - XB_{*k})^T(y_k - XB_{*k}); \tag{2}$$

The linear regression model may overfit especially when $P \gg N$ since X^TX will be singular. Regularizations in terms of penalties [18], [26], [29], [30] are introduced to avoid overfitting, whereas the optimization problem can be defined as

$$\begin{aligned}\hat{B}_{*k} = argmin_{B_{*k}}\ & \frac{1}{2}(y_k - XB_{*k})^T(y_k - XB_{*k}) \\ & +Penalty_k, \forall k = 1,...,Q.\end{aligned} \tag{3}$$

Usually only a few input features are related to each output feature so as to avoid overfitting. In order to minimize the number of non-zero entries in B_{*k}, an intuitive penalty should be $\lambda_k||B_{*k}||_0$ (count of non-zero elements in B_{*k}) with larger $\lambda_k(\lambda_k > 0)$, but unfortunately this induces to an NP-hard problem. By setting $Penalty_k = \lambda_k||B_{*k}||_1(\lambda_k > 0)$ [26], the problem can be solved efficiently by the least angle regression [12] or coordinate descent algorithm [28]. Ridge regression sets $Penalty_k = \lambda_k||B_{*k}||_2$ to shrink the B_{*k} coefficients to be small. Many studies [18] have shown that Ridge can reduce overfitting when there are many highly related input features. By using a mixture penalty of norm one and norm two, Elastic Net [30] sets $Penalty_k = \lambda_k(\alpha_k||B_{*k}||_1 + \frac{1-\alpha_k}{2}||B_{*k}||_2^2)$ where $\lambda_k > 0$ and $0 \le \alpha_k \le 1$, which becomes Lasso or Ridge when $\alpha_k = 1$ or $\alpha_k = 0$ respectively. Since finding the optimal B_{*k} in (2) is equivalent to find the best B_{*k} such that y_k follows the Gaussian distribution $\mathcal{N}(XB_{*k}, \sigma^2 I_N)$ where σ^2 is variance and I_N is identity matrix. The models with different kinds of regularizations in (3) can be viewed as finding maximum likelihood B_{*k} under different assumptions of the distributions of entries in B_{*k}. Lasso, Ridge and Elastic Net can be treated as MAPs (Maximum a posterior) of Laplacian, Gaussian and mixture of Laplacian and Gaussian distribution on B_{*k}. Even these methods for single output features could be easily extended to multiple output features cases by considering each output feature independently, studies [20], [22], [24] show that by considering multiple output features simultaneously, the estimation of regression coefficients may benefit from taking into account the underlying relationships shared by different outputs features. Similarly, for multiple output setting we can model the relationship between input features and output features as multiple output regression and the parameters can be estimated as follows:

$$\hat{B} = argmin_B \sum_{k=1}^{Q} \frac{1}{2}(y_k - XB_{*k})^T(y_k - XB_{*k}) + Penalty. \tag{4}$$

Multi-Lasso, Multi-Ridge and Multi-Elastic Net set *Penalty* to be $\lambda\sum_{j=1}^{P}||B_{j*}||_1$, $\lambda\sum_{j=1}^{P}||B_{j*}||_2^2$ and $\lambda\sum_{j=1}^{P}(\alpha||B_{j*}||_1 + \frac{1-\alpha}{2}||B_{j*}||_2^2)$ respectively.

A. Group Sparsity

We learn the interaction effects (each row of B) of input features to all output features together. Multi-Lasso assumes that each row of matrix B should have similar level of sparsity and tends to penalize each row of matrix B with an uniform penalty. It does not consider the case that some input features are related to less output features while some are related to more. In reality it is reasonable to assume that different rows have different sparsity structures and different sparsity level dependent penalties should be considered. We consider a two-level framework for model group sparsity. Firstly, when considering intergroup (different rows of matrix B) sparsity, we assume that only a small number of input features may be related to the output features, i.e. many rows of matrix B will be zero. Secondly, for intragroup (each row of matrix B) sparsity, we assume that each row of matrix B will also be sparse, but has different level of intragroup sparsity. We set $Penalty = \lambda\sum_{j=1}^{P} a_j||B_{j*}||_1$ where $\lambda > 0$, and $a_j > 0\ \forall j = 1,...,P$. We define $A = [a_1,...,a_P]^T$ to model the intragroup sparsity where a_j is non-negative scaling parameter for modelling the sparsity level for the j^{th} row of matrix B. Intuitively larger a_j will induce more sparse solutions on j^{th} row of B, and smaller a_j will have more non-zero entries. When the dimensionality of input features is high (number of rows of matrix B is also large), it is impractical to determine the regularization parameters a_j by cross validation while our method can automatically choose those a_js which can best fit the data. The penalty λ is used to model the intergroup sparsity. Larger λ value means more rows of matrix B to be zero while smaller λ has less. We model the two-level sparsity pattern by assuming different distributions on B and propose the following model:

$$\hat{A}, \hat{B} = argmin_{A,B} \sum_{k=1}^{Q} \frac{1}{2}(y_k - XB_{*k})^T(y_k - XB_{*k}) + Penalty. \tag{5}$$

By considering a serial of Laplacian distributions on B, we propose the penalty for LapMOR (**Lap**lacian **M**ultiple **O**utput **R**egression) which is defined as follows:

$$\begin{aligned} Penalty &= -log\Bigg(\prod_{j=1}^{P}\prod_{k=1}^{Q}\frac{\lambda a_j}{2}\exp^{-\lambda a_j|B_{jk}|}\Bigg) \\ &= \lambda\sum_{j=1}^{P} a_j||B_{j*}||_1 - Q\sum_{j=1}^{P} log\ a_j + const. \\ & s.t.\ \lambda > 0,\ a_j > 0\ \forall j = 1,...,P \\ & \sum_{j=1}^{P} a_j = 1. \end{aligned} \tag{6}$$

The first term in (5) is the error for fitting and (5) is the regularization term with a_j automatically learned from data. By setting $\sum_{j=1}^{P} a_j = 1$, the learned optimal a_j can be interpreted as the relative sparsity for j^{th} row with larger a_j for more sparse B_{j*}. In order to reduce the unstable estimations made by LapMOR introduced by the highly dependent input features, we further extend the model by imposing a serial of mixtures of Laplacian and Gaussian distributions on B_{j*} and propose the following penalty for model LGMOR (**L**aplacian and **G**aussian **M**ultiple **O**utput **R**egression):

$$\begin{aligned} Penalty = -log\Bigg(& (\prod_{j=1}^{P}\prod_{k=1}^{Q}\frac{\lambda\alpha a_j}{2}\exp^{-\lambda\alpha a_j|B_{jk}|}) \\ & \bullet (\prod_{j=1}^{P}\prod_{k=1}^{Q}\frac{\sqrt{\lambda(1-\alpha)a_j}}{\sqrt{2\pi}}\exp^{-\frac{\lambda(1-\alpha)a_j}{2}B_{jk}^2})\Bigg) \\ = & \lambda\alpha\sum_{j=1}^{P} a_j||B_{j*}||_1 + \frac{\lambda(1-\alpha)}{2}\sum_{j=1}^{P} a_j||B_{j*}||_2^2 \\ & -\frac{3Q}{2}\sum_{j=1}^{P} log\ a_j + const, \\ & s.t.\ \lambda > 0,\ 0 \le \alpha \le 1,\ a_j > 0\ \forall j = 1,...,P \\ & \sum_{j=1}^{P} a_j = 1. \end{aligned} \tag{7}$$

LGMOR uses a more general assumption (mixture of Laplacian and Gaussian) on the distribution of regression coefficients rather than Laplacian alone. It takes advantage of both Laplacian distribution for sparsity and Gaussian distribution for shrinkage. There is always a trade-off between the sparsity level introduced by Laplacian and the shrinkage estimation from the Gaussian.

We use cross validation to choose λ, α in LGMOR by choosing the parameter with best prediction performance. The prediction performances on normalized validation datasets with various αs and λs are shown in Fig. 1 (for family cut dataset in section III).

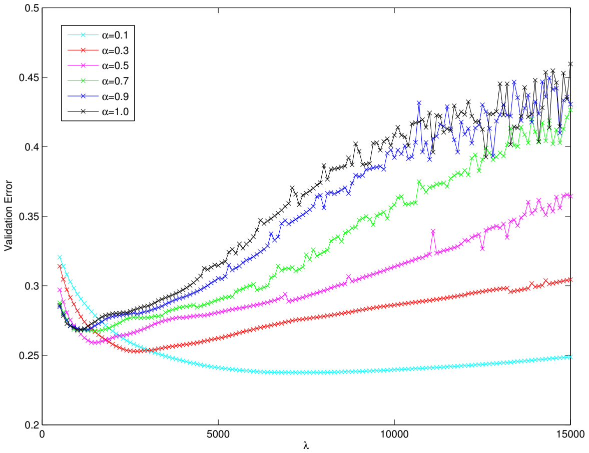

Fig. 1. The validation RMSE by using different λs and αs for LGMOR

B. Parameter Estimation

Although the optimization problem in equation (5) is non-convex, by fixing either A or B, the equation (5) becomes convex for B or A with respect to the other parameter, which can be solved efficiently by alternatively solving the equation with A or B fixed [21] in Algorithm 1. By fixing A, we can solve (5) efficiently by least angle regression or coordinate descent algorithm. By fixing B, we can solve (5) by the projected gradient descent method [11].

III. Empirical Studies

We compare the performance of our approaches LapMOR and LGMOR with other approaches linear regression, independent Lasso, independent Elastic Net, dirty model, rMTFL, LapMOR and LGMOR on both synthetic and real datasets. Experiments show that our method can reduce the false positive rates and produce better prediction results on the testing datasets. We use the packages glmnet [13], [14] and MALSAR [6] for Lasso, Elastic net and Dirty model, rMTFL for comparison.

A. Synthetic data

We generate data with $N = 60$ samples, $P = 100$ input features and $Q = 10$ output features. The matrix X is generated with each feature sampled randomly and independently from Gaussian distribution $\mathcal{N}(0,1)$. We set 65 rows of matrix B to be exactly zero. For the remaining 35 rows, we divide them into five groups, each with seven rows. Each element in the i^{th} group is set to be non-zero with probability $\frac{i}{5}$ for $1 \le i \le 5$. All non-zero elements in B are randomly generated with Gaussian distribution $\mathcal{N}(0,0.5)$, and two Gaussian noises ε_1 and ε_2 with distribution $\mathcal{N}(0,0.02)$. Finally we set $Y = XB + \varepsilon_2$ and update $X = X + \varepsilon_1$. By using cross validation on all 60 samples (each time 48 samples for training and the left 12 for validation), we set the optimal λ for LapMOR and LGMOR ($\alpha = 0.5$) to be 21 and 22 respectively. For easy comparison of the performance, we fix $\alpha = 0.5$ for both Elastic Net and LGMOR. The plots of generated and estimated parameters B are shown in Fig. 2. The receiver operating characteristic (ROC) curves of various methods for Bs predicted by different methods are shown in Fig. 3 (darker grids mean larger absolute values). From the ROC curve, LapMOR and LGMOR have better performances with less False Negatives and True Negatives. Lasso performs better than Elastic Net because the simulated data does not contain any highly dependent input features, similarly for LapMOR and LGMOR. Dirty model performs

Algorithm 1: Algorithm for solving equation (5)

```
Input:
    X: an $N \times P$ matrix for input features;
    Y: an $N \times Q$ matrix for output features;
    $\alpha$: percentage for Laplacian prior (for LGMOR only);
    $\lambda$: penalizing coefficient;
    $\varepsilon$: convergence threshold.
Output:
    A: a length $P$ vector $[a_1,...,a_P]$ for different rows of matrix $B$;
    $B$: a $P \times Q$ regression coefficient $[B_{jk}]_{P \times Q}$ matrix.
for $j = 1$ to $P$ do
    $a_j^{(0)} = \frac{1}{P}$;
    for $k = 1$ to $Q$ do
        $B_{jk}^{(0)} = Random\ Number$;
$t = 0$;
repeat
    $t = t + 1$;
    Case LapMOR:
    $B^{(t)} = argmin_B \sum_{k=1}^{Q} \frac{1}{2}(y_k - XB_{*k})^T(y_k - XB_{*k}) + \lambda \sum_{j=1}^{P} a_j^{(t-1)} ||B_{j*}||_1$;
    $A^{(t)} = argmin_A \ \lambda \sum_{j=1}^{P} a_j ||B_{j*}^{(t)}||_1 - Q \sum_{j=1}^{P} \log a_j$, such that $a_j > 0 \ \forall j = 1,...,P$ and $\sum_{j=1}^{P} a_j = 1$;
    Case LGMOR:
    $B^{(t)} = argmin_B \sum_{k=1}^{Q} \frac{1}{2}(y_k - XB_{*k})^T(y_k - XB_{*k}) + \lambda\alpha \sum_{j=1}^{P} a_j^{(t-1)} ||B_{j*}||_1 + \frac{\lambda(1-\alpha)}{2} \sum_{j=1}^{P} a_j^{(t-1)} ||B_{j*}||_2^2$;
    $A^{(t)} = argmin_A \ \lambda\alpha \sum_{j=1}^{P} a_j ||B_{j*}^{(t)}||_1 + \frac{\lambda(1-\alpha)}{2} \sum_{j=1}^{P} a_j ||B_{j*}^{(t)}||_2^2 - \frac{3Q}{2} \sum_{j=1}^{P} \log a_j$, such that $a_j > 0 \ \forall j = 1,...,P$ and $\sum_{j=1}^{P} a_j = 1$;
until $||A^{(t)} - A^{(t-1)}||_\infty < \varepsilon \ and \ ||B^{(t)} - B^{(t-1)}||_\infty < \varepsilon$;
$A = A^{(t)}, B = B^{(t)}$;
```

poorly because the sparsity for each row of matrix B is very strict since without considering the outliers. Note that these methods consider an uniform penalty parameter which leads to their poor performances.

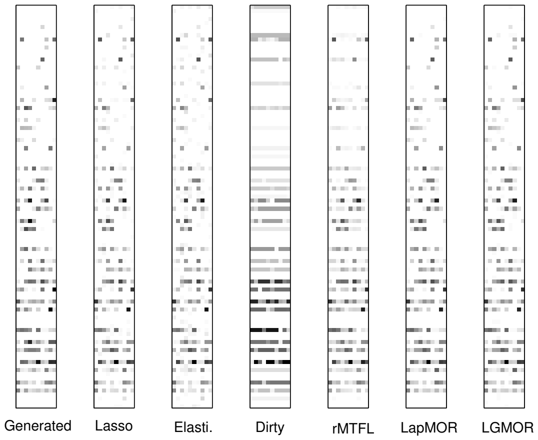

Fig. 2. Real parameters, Lasso, Elastic Net ($\alpha = 0.5$), Dirty Model, rMTFL, LapMOR and LGMOR ($\alpha = 0.5$)

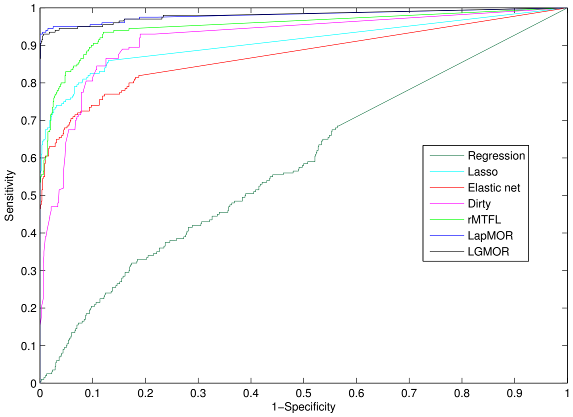

Fig. 3. ROC curve for Regression, Lasso, Elastic Net ($\alpha = 0.5$), Dirty Model, rMTFL, LapMOR, LGMOR ($\alpha = 0.5$)

B. *Real World Data*

We use the published sequencing data from the ICoMM (International Census of Marine Microbes) project [25] (available in http://icomm.mbl.edu/microbis/) for comparison. The ICoMM database contains altogether 644 environmental samples collected from sea water, river water, sediments or biofilms in 297 different geography sites over the world. The published sequencing data includes sequences from the bacteria communities of 487 samples (in 246 sites), which are sequenced from the V6 region of the prokaryote 16S rRNA. The sequenced data is processed by the VAMPS pipeline (http://vamps.mbl.edu) and altogether 8,570, 814 cleaned, annotated sequences (1,378,983 non-redundant) are published.

1) Preprocessing of ICoMM Data: We employ ESPIRIT-Tree [8] to cluster sequences into OTUs at various distance levels. To tackle the problem that ESPRIT-Tree uses a fixed distance threshold to define OTUs (operational taxonomy units) [7], which is inconsistent with genomic variations between taxa in real-world. We further process the clustering tree obtained by ESPRIT-Tree using a semi-supervised method called VI-Cut [23]. VI-Cut adopts the portion of annotated sequences in the dataset to aid the partitioning of OTUs, which generalizes well on unknown groups [27]. It should be noted that the annotations given by VAMPS may contain errors, which will degrade the quality of OTU picking. Hence, before running VI-Cut, we examine the OTUs generated by ESPRIT-Tree at distance levels, which are believed to have no family and genus mixing. If sequences from two or more family- or genus-levels co-exist in the same OTU, the annotations are considered unreliable. As a result, 18,620 family-level OTUs and 26,185 genus-level OTUs, are extracted using the above pipeline.

2) Experiment Results: We select the top 500 abundant OTUs for family and genus cut with abundance of 91.3% and 86.07% among all samples. We select 271 samples with complete depth, temperature and salinity information, these three factors, chosen based on principal coordinate analysis, are strongly related to most variant principal coordinates (PCs) and consistent with previous findings such as [16]. We extract the OTU abundance matrices from the top 500 abundant OTUs for both family and genus cut OTUs with depth, temperature and salinity information to form the environmental factor matrix, i.e. OTU abundance matrix as input feature matrix X and environmental matrix as output feature matrix Y. We get the value ranges for depth (between 0.000 and 11.949 on log scale), salinity (between 0.000 and 40.878), and temperature (between -2.000 and 28.700). The OTUs and environmental factors are normalized to have zero mean and unit variance. The data are divided into 200 samples for training and the remaining 71 samples for testing. For obtaining the optimal λ and α, we use five-fold cross validation. We repeat the experiment for 66 times by randomly partitioning our dataset. The results of average RMSE (root mean square error) and derivative for family cut dataset and genus cut dataset are shown in Table I and Table II respectively. By using five-fold cross validation, for LapMOR, we set $\lambda = 1000$ for both family cut and genus cut dataset, similarly, $\alpha = 0.1$ and $\lambda = 7000$ for LGMOR for both datasets. Regression has the worst performance since the empirical covariance $X^T X$ is singular which leads to overfitting. LapMOR and LGMOR perform better on the OTU datasets than others with lower variances of prediction errors. Elastic Net performs better than Lasso since there are a lot of highly dependent OTUs, similar reason for the performance of LapMOR and LGMOR. rMTFL performs better than dirty model since it considers a looser restriction on the row sparsity. As our method is more flexible on the rows sparsity and shrinkage estimation, our methods have the best performance and they fit better on the training dataset than others too except for regression.

IV. CONCLUSIONS

In this paper, we propose a multiple output regression method with intra- and inter-group sparsity and shrinkage estimation. We propose a probabilistic formulation for the

TABLE I. PREDICTION PERFORMANCE COMPARISON FOR FAMILY CUT DATASET

	Depth	Salinity	Temperature	Average
Regre.	8.298±2.063	11.091±2.411	9.467±2.629	9.618±2.368
Lasso	2.195±0.584	4.711±1.646	4.931±3.957	3.946±2.062
Elast.	1.905±0.428	3.845±0.970	5.149±4.864	3.633±2.087
Dirty	2.263±0.226	4.730±1.000	4.498±0.535	3.830±0.587
rMTFL	1.860±0.864	4.013±1.730	4.048±1.770	3.307±1.454
LapMOR	1.674±0.174	3.878±0.974	3.084±0.501	2.879±0.550
LGMOR	**1.513±0.152**	**3.533±0.800**	**2.967±0.527**	**2.671±0.493**

TABLE II. PREDICTION PERFORMANCE COMPARISON FOR GENUS CUT DATASET

	Depth	Salinity	Temperature	Average
Regre.	6.556±1.857	8.250±2.100	8.864±2.271	7.889±2.076
Lasso	1.962±0.331	4.229±1.205	4.377±2.184	3.523±1.240
Elast.	1.798±0.325	3.960±1.329	3.746±0.921	3.169±0.858
Dirty	1.948±0.225	4.092±0.907	3.924±0.654	3.321±0.595
rMTFL	1.785±0.603	3.908±0.976	3.781±0.946	3.158±0.842
LapMOR	1.535±0.174	3.706±0.838	3.038±0.499	2.760±0.503
LGMOR	**1.456±0.135**	**3.426±0.728**	**2.993±0.526**	**2.625±0.463**

distributions of regression coefficients (Mixture of Laplacian and Gaussian distributions). We apply our method to the global ICoMM dataset to study the interaction effects between microbe communities and environmental factors. Our method outperforms other methods such as Lasso, Elastic Net, dirty model and rMTFL on prediction.

V. ACKNOWLEDGMENTS

This work was supported by Shenzhen basic research project (NO.JCYJ20120618143038947).

REFERENCES

[1] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, Anne U Jackson, et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345, 2007.

[2] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[3] Jianhui Chen, Ji Liu, and Jieping Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 5(4):22, 2012.

[4] Bin Cheng, Guangcan Liu, Jingdong Wang, Zhongyang Huang, and Shuicheng Yan. Multi-task low-rank affinity pursuit for image segmentation. In *2011 IEEE International Conference on Computer Vision (ICCV)* , pages 2439–2446. IEEE, 2011.

[5] Seunghak Lee, Jun Zhu, and Eric P Xing. Adaptive multi-task lasso: with application to eqtl detection. In *Advances in Neural Information Processing Systems*, pages 1306–1314, 2010.

[6] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Malsar: Multi-task learning via structural regularization. *Arizona State University*, 2012.

[7] Mark Blaxter, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd, and Eyuanlem Abebe. Defining operational taxonomic units using dna barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:55–67, 2005.

[8] Yunpeng Cai and Yijun Sun. Esprit-tree: hierarchical clustering analysis of millions of 16s rrna pyrosequences in quasilinear computational time. *Nucleic Acids Research*, 39(14):e95–e95, 2011.

[9] Xiaohui Chen, Xinghua Shi, Xing Xu, Zhiyong Wang, Ryan Mills, Charles Lee, and Jinbo Xu. A two-graph guided multi-task lasso approach for eqtl mapping. In *International Conference on Artificial Intelligence and Statistics*, pages 208–217, 2012.

[10] Rebecca W Doerge. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, 3(1):43–52, 2002.

[11] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine learning*, pages 272–279. ACM, 2008.

[12] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004.

[13] Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[14] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[15] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

[16] Pierre E Galand, Emilio O Casamayor, David L Kirchman, and Connie Lovejoy. Ecology of the rare microbial biosphere of the arctic ocean. *Proceedings of the National Academy of Sciences*, 106(52):22427–22432, 2009.

[17] Pinghua Gong, Jieping Ye, and Changshui Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–903. ACM, 2012.

[18] ARTHUR E. Hoerl and ROBERT W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:301–320, 1970.

[19] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 964–972, 2010.

[20] Seyoung Kim and Eric P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of International Conference on Machine Learning*, pages 543–550, 2010.

[21] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2006.

[22] Aurelie C. Lozano and Grzegorz Swirszcz. Multi-level lasso for sparse multi-task regression. In *Proceedings of International Conference on Machine Learning*, 2012.

[23] Saket Navlakha, James White, Niranjan Nagarajan, Mihai Pop, and Carl Kingsford. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. In *Research in Computational Molecular Biology*, pages 400–417. Springer, 2009.

[24] Guillaumi Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. Technical report, Department of Statistics, University of California, Berkeley.

[25] Mitchell L Sogin, Hilary G Morrison, Julie A Huber, David Mark Welch, Susan M Huse, Phillip R Neal, Jesus M Arrieta, and Gerhard J Herndl. Microbial diversity in the deep sea and the underexplored ”rare biosphere”. *Proceedings of the National Academy of Sciences*, 103(32):12115–12120, 2006.

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

[27] James R White, Saket Navlakha, Niranjan Nagarajan, Mohammad-Reza Ghodsi, Carl Kingsford, and Mihai Pop. Alignment and clustering of phylogenetic markers-implications for microbial diversity studies. *BMC Bioinformatics*, 11(1):152, 2010.

[28] Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2:224–244, 2008.

[29] Ming Yuan. Model selection and estimation in regression with group variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2006.

[30] Hui Zhou and Trevor Hastie. Regularization and varible selection via the elastic net. *Journal of the Royal Statistical Society*, 67:301–320, 2005.