JMLR: Workshop and Conference Proceedings 30:1–15, 2015 ACML 2015

Robust Multivariate Regression with Grossly Corrupted Observations and Its Application to Personality Prediction

Xiaowei Zhang ZHANGXW@BII.A-STAR.EDU.SG
Li Cheng CHENGLI@BII.A-STAR.EDU.SG
*Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore*

Tingshao Zhu TSZHU@PSYCH.AC.CN
Institute of Psychology, Chinese Academy of Sciences, China

Abstract

We consider the problem of multivariate linear regression with a small fraction of the responses being missing and grossly corrupted, where the magnitudes and locations of such occurrences are not known in priori. This is addressed in our approach by explicitly taking into account the error source and its sparseness nature. Moreover, our approach allows each regression task to possess its distinct noise level. We also propose a new algorithm that is theoretically shown to always converge to the optimal solution of its induced non-smooth optimization problem. Experiments on controlled simulations suggest the competitiveness of our algorithm comparing to existing multivariate regression models. In particular, we apply our model to predict the *Big-Five* personality from user behaviors at Social Network Sites (SNSs) and microblogs, an important yet difficult problem in psychology, where empirical results demonstrate its superior performance with respect to related learning methods.

1. Introduction

It is known that multivariate linear model (Anderson, 2003) takes the form

$$Y = XW^* + Z, \tag{1}$$

where $Y \in \mathbb{R}^{n\times p}$ is a response matrix, $X \in \mathbb{R}^{n\times d}$ is a design matrix, $W^* \in \mathbb{R}^{d\times p}$ is an unknown regression coefficient matrix and $Z \in \mathbb{R}^{n\times p}$ is a stochastic observation noise matrix. The central problem then is to accurately estimate the coefficient regression matrix W^* from noisy observations Y. Typically the noise is assumed to have bounded energy and can be successfully absorbed into the noise matrix Z, which is usually modelled as following Gaussian-type distributions. As is standard in statistics, a general framework of estimating W^* is given by

$$\hat{W} = \arg\min_{W}\{\mathcal{L}(X, Y; W) + \lambda R_W(W)\}, \tag{2}$$

where $\mathcal{L}$ is a loss function, $\lambda > 0$ is a user-specified tuning parameter and $R_W(W)$ is some regularization function. One of the most popular choices for $\mathcal{L}$ is the least square loss which achieves the optimal rates of convergence under some conditions on X and Z (Lounici et al., 2011; Rohde and Tsybakov, 2011) and has found many applications such as multi-task learning (Caruana, 1997; Argyriou et al., 2008).

However, there are practical scenarios where a few data entries are corrupted by severe noises who are considerably larger than the "normal" ones that can be incorporated in above-mentioned noise model. For example, there could be careless or even malicious user annotations that severely corrupt a few data entries, and yet are difficult to be spotted. This type of noises might dramatically degrade the estimation accuracy of estimator $\hat{W}$.

In this paper, we propose to address this gross corruption problem by *explicitly* introducing a sparse matrix $G^* \in \mathbb{R}^{n\times p}$, where the locations of nonzero entries are unknown and their magnitudes could be very large. This gives the following linear model

$$Y = XW^* + Z + G^*. \tag{3}$$

It thus provide us the ability to recover the grossly corrupted examples instead of simply trashing them away as outliers. The corresponding optimization framework is naturally extended to that of estimating (W^*, G^*) by solving

$$\min_{W,G}\{\overline{\mathcal{L}}(X, Y; W, G) + \lambda R_W(W) + \rho R_G(G)\}, \tag{4}$$

with $\rho > 0$ being a tuning parameter and $R_G(G)$ being some regularization function. In addition, instead of the usual least square loss, the $\ell_{2,1}$-norm is adopted as the loss function to better cope with noise. Moreover, a group sparsity inducing norm is used for $R_W(W)$ to enforce group-structured sparsity, as well as ℓ_1-norm for $R_G(G)$ to impose element-wise sparsity on identifying possible gross corruptions. The main contributions here are threefold: (1) Our approach explicitly model and recover the grossly corrupted examples in the context of multivariate linear regression with the ℓ_1-norm as loss function; (2) An efficient and convergent multi-block proximal alternating direction method of multipliers (ADMM) is applied to the induced non-smooth optimization problem; (3) In particular, we focus on the application of Big-Five personality prediction based on behaviors at social network sites (SNSs), a novel and important problem in computational psychology. Experiments with synthetic and real data demonstrate the advantages of our method with respect to existing multivariate regression models.

Related Work in Machine Learning and Statistics For linear regression model under gross error as in (3), there have been various lines of work (Li, 2012; Nguyen and Tran, 2013; Wright and Ma, 2010; Xu et al., 2013, 2012; Xu and Leng, 2012), among which (Nguyen and Tran, 2013) and (Xu and Leng, 2012) consider the same optimization problem as (4) for univariate and multivariate regression, respectively. However, both of them adopt least square loss which, as pointed out in (Liu et al., 2013, 2014), has two drawbacks: First, all regression tasks are regularized by the same parameter λ, which ignores the disparate noise levels contained in different regression tasks; Second, the selection of the optimal λ is dependent on the estimation of unknown variance of Z, which is crucial for achieving a good finite-sample performance. To overcome these two drawbacks, a calibrated multivariate regression (CMR) method has been proposed in (Liu et al., 2013, 2014), which uses the $\ell_{2,1}$-norm as its loss function. It is tuning insensitive and calibrates each task with respect to its own noise level, and it has been shown both theoretically and empirically that CMR achieves an improved finite sample performance. As our approach also adopts the $\ell_{2,1}$-norm as the loss function as in (Liu et al., 2013, 2014). In this sense, our model in (4) may

be considered as an extension of the CMR model to deal with the challenging problem of regression with grossly corrupted observations. It is worth noting that our induced optimization problem and thus the solver are clearly different from those of (Liu et al., 2013, 2014).

Related Work in Personality Prediction In psychology, the most influential theory in the study of personality is the so called Big-Five theory (Funder, 2001; Matthews et al., 2006), where personality is characterized by five distinct dimensions: *Agreeableness*, *Conscientiousness*, *Extraversion*, *Neuroticism* and *Openness*. To identify an individual's personality, the most commonly used form is the *self-report inventory* (Domino and Domino, 2006), which requires people to answer questions about their typical behavior. In addition, the Berkeley Personality Lab [1] has designed a widely-used Big-Five Inventory (BFI), which contains 44 questions with high validity and reliability, and can be further summarized to form a quantized five-dimensional personality descriptor. Despite its wide usage, it is practically very difficult to conduct self-report inventory in large scales. There are also issues related to careless or malicious user annotations.

Since personality can be manifested by behavior, a natural idea is to infer personality from human behaviour. Over the past decade, online Social Network Sites (SNSs) such as Facebook and Twitter, as well as their Chinese equivalent, RenRen and Sina Weibo, have become popular means of social communication and have been regarded as an important part of people's daily life. It has been observed that online SNS behaviors and real-world behaviors share a significant amount in common (Landers and Lounsbury, 2006). SNSs also provide an excellent data source for social research, since they are able to reproduce social activities in real life and acquire rich information derived from a large and diverse population. The growing demand of SNS users facilitates the technical development of personalized recommendation (Jie, 2011; Reynol, 2011). One major issue here is to design a proper personality prediction model.

So far, there are only sporadic research efforts that relate personality and SNS behaviors (Ma et al., 2011). Orr et al. (Orr et al., 2009) discuss the influence of shyness on the use of SNS and discover that shyness is significantly positively associated with the time spent on SNS and negatively correlated with the number of "friends". Correa et al. (Correa et al., 2010) analyze the connection of users' personality and social media and find that openness and extraversion associate positively with social media usage while neuroticism associates negatively. Nevertheless, these works could only provide the association relation between personality and behavior instead of a direct and quantitative measurement.

The closest research effort is probably that of Gosling et al. (Gosling et al., 2011), where experiments are conducted towards the manifestations of personality in SNS, to deliver a mapping between personality and SNS online behaviors. They examine the personality with *self-reported* Facebook usage and observable profile information and provide the correlation factor between personality and online behaviors. In their work eleven features are used including friends count, weekly usage and other frequency-related features. However, these features are all based on statistical characteristics instead of the inner properties of users. Moreover, their data collection procedure relies on self-reported usage and observable profile

1. http://www.ocf.berkeley.edu/~johnlab/index.html

information – both require considerable manual efforts and are not realistic for practical purpose.

Notations We summarize here some useful notations used throughout this paper. For any scale α, $(\alpha)_+ := \max\{\alpha, 0\}$. For any $\boldsymbol{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$ and $1 \leq p < \infty$, we denote by $\|\boldsymbol{x}\|_p := (\sum_{i=1}^d |x_i|^p)^{\frac{1}{p}}$ the ℓ_p norm and $\|\boldsymbol{x}\|_\infty := \max_{i=1}^d |x_i|^p$ the ℓ_∞ norm. A group g is a subset of $\{1, \cdots, d\}$ with cardinality $|g|$, while $\mathcal{G}$ denotes a set of potentially overlapping groups such that $\cup_{g\in\mathcal{G}} g = \{1, \cdots, d\}$. $\boldsymbol{x}_g$ denotes the subspace of $\boldsymbol{x} \in \mathbb{R}^d$ with dimensions indexed by g. Similarly, for any matrix $A \in \mathbb{R}^{d\times n}$, we denote by A_{g*} and A_{*g} the rows and the columns of A indexed by g, respectively. I denotes the identity matrix whose size should be clear from the context. For any matrix A, we respectively define the spectral, Frobenius and ℓ_∞ norms as: $\|A\|_2 := \max_{1\leq i\leq r} \sigma_i(A)$, $\|A\|_F := \sqrt{\sum_{ij} A_{ij}^2}$ and $\|A\|_\infty := \max_{ij} |A_{ij}|$, where r is the rank of A and $\sigma_i(A)$ is the i-th largest singular value of A. The following matrix norms are also specified: $\|A\|_{1,2} := \sum_{i=1}^d \|A_{i*}\|_2$, $\|A\|_{2,1} := \sum_{j=1}^n \|A_{*j}\|_2$. Finally, given a group set $\mathcal{G}$, we denote by $\Omega_{\mathcal{G}}(A) := \sum_{g\in\mathcal{G}} \|A_{g*}\|_F$ the group lasso penalty associated with $\mathcal{G}$.

2. Our Model

Throughout this paper, we assume that $R_W(W) = \Omega_{\mathcal{G}}(W)$ is a group sparsity inducing norm and $Z_{i*} \overset{\text{iid}}{\sim} N(0, \Sigma)$ for some covariance matrix Σ.

The standard multivariate regression model with least square loss considers the following convex problem

$$\hat{W} = \arg\min_W \frac{1}{\sqrt{n}} \|Y - XW\|_F^2 + \lambda\Omega_{\mathcal{G}}(W). \tag{5}$$

Let $\sigma_{max} = \max_{1\leq k\leq p} \sigma_k$, it has been shown in (Lounici et al., 2011) that, under the assumption that

$$\Sigma = \text{diag}(\sigma_1^2, \cdots, \sigma_p^2), \tag{6}$$

and suitable conditions on X, if we choose $\lambda = 2c\sigma_{max}(\sqrt{\ln d} + \sqrt{p})$ for some $c > 1$, then the estimator $\hat{W}$ in (5) achieves the optimal rates of convergence. In (Liu et al., 2013, 2014), a new method named CMR has been proposed that calibrates the regularization for each regression task with respect to its noise level σ_k by considering the following optimization problem:

$$\hat{W} = \arg\min_W \|Y - XW\|_{2,1} + \lambda\Omega_{\mathcal{G}}(W). \tag{7}$$

The optimal rate convergence of CMR in parameter estimation has also been provided in (Liu et al., 2013, 2014).

Notice that all previous work fails to take into considerations the gross errors such as missing and grossly corrupted responses. To handle this kind of problem, it is natural to consider the model (3). In particular, we attempt to combine the advantages of the CMR method (7) and the model in (3) by considering the following problem:

$$(\hat{W}, \hat{G}) = \arg\min_{W,G} \|Y - XW - G\|_{2,1} + \lambda\Omega_{\mathcal{G}}(W) + \rho\|G\|_1, \tag{8}$$

where $\lambda > 0$ and $\rho > 0$ are tuning parameters. Our method can be considered as an extension of CMR in the sense that when there is no gross error, our method becomes the original CMR.

A related work is (Xu and Leng, 2012) that considers the same model as (3) and studies robust multi-task regression with grossly corrupted observations by solving the following optimization problem:

$$(\hat{W}, \hat{G}) = \arg\min_{W,G} \frac{1}{2}\|Y - XW - G\|_F^2 + \lambda\|W\|_{1,2} + \rho\|G\|_1. \tag{9}$$

However, as pointed out in (Liu et al., 2013, 2014), there are some limitations in the least square loss function in (9).

In the context of Big-Five personality prediction, the response is a five dimensional vector for each observed instance (i.e., a session of one's SNS behaviors). It is clear that these five dimensions are distinct and yet *intrinsically connected*: For example, sociable individuals (i.e. Extraversion) are often more likely to be cooperative (i.e. Agreeableness). Moreover, the noise levels of these five regression tasks are not necessarily the same. This naturally suggests the usage of Multi-Task regression model with CMR loss. On the other hand, ground-truth responses are obtained based on self-report inventory. It is subjective by nature, and is truly difficult to rule out the potential existence of gross errors from either careless or malicious user annotations.

3. Our *CMRG* Algorithm

Since both the loss function and regularization terms are non-smooth, the optimization problem in (8) is computationally more difficult to solve than standard multivariate linear regression problems in (5) and (9). In this section, we apply the proximal ADMM (Fazel et al., 2013; Sun et al., 2014) to solve the optimization problem in (8). To start with, let us reformulate (8) as an equivalent linearly constrained problem, using the variable splitting procedure inspired by (Chen et al., 2012; Qin and Goldfarb, 2012).

Denote $d' := \sum_{g\in\mathcal{G}} |g|$, and let $\Theta \in \mathbb{R}^{d'\times p}$ be the matrix obtained from W by duplicating those rows of W whenever they are shared between two overlapping groups. In other words, for the set of overlapping groups $\mathcal{G}$, there exists a set of groups $\mathcal{G}'$ consisting of non-overlapping groups, which forms a disjoint partition of $\{1, \cdots, d'\}$ such that the following identity holds

$$\Omega_{\mathcal{G}}(W) = \sum_{g\in\mathcal{G}} \|W_{g*}\|_F = \sum_{g'\in\mathcal{G}'} \|\Theta_{g'*}\|_F = \Omega_{\mathcal{G}'}(\Theta).$$

Clearly $d' \geq d$, and the relationship between W and Z is specified by the linear system $CW = Z$, where $C \in \mathbb{R}^{d'\times d}$ is defined as: $C_{ij} = 1$ if $Z_{i*} = W_{j*}$ and $C_{ij} = 0$ otherwise. It is easy to see that C is a highly sparse matrix, and $D := C^\top C$ is a diagonal matrix with each diagonal entry equalling the number of repetitions of the corresponding row of W. When $\mathcal{G}$ consists of non-overlapping groups, we have $C = I$.

Let $Z = Y - XW - G$, the optimization problem in (8) can be equivalently reformulated as

$$(\hat{W}, \hat{G}) = \arg \min_{(Z,\Theta),G,W} \|Z\|_{2,1} + \lambda \Omega_{\mathcal{G}'}(\Theta) + \rho \|G\|_1 \\ s.t. \begin{bmatrix} Z \\ \Theta \end{bmatrix} + \begin{bmatrix} I_n \\ 0 \end{bmatrix} G + \begin{bmatrix} X \\ -C \end{bmatrix} W = \begin{bmatrix} Y \\ 0 \end{bmatrix}, \tag{10}$$

which is in the exact form of the following 3-block convex optimization problem:

$$\min\{f(\boldsymbol{y}) + g(\boldsymbol{z}) - \langle \boldsymbol{b}, \boldsymbol{w} \rangle \mid \mathcal{F}\boldsymbol{y} + \mathcal{H}\boldsymbol{z} + \mathcal{B}\boldsymbol{w} = \boldsymbol{c}\}, \tag{11}$$

if we let $\boldsymbol{y} := [Z^\top \ \Theta^\top]^\top$, $\boldsymbol{z} := G$, $\boldsymbol{w} := W$, $f(\boldsymbol{y}) := \|Z\|_{2,1} + \lambda \Omega_{\mathcal{G}'}(\Theta)$, $g(\boldsymbol{z}) := \rho \|G\|_1$ and $\boldsymbol{b} = 0$.

To solve the above 3-block convex optimization problem, one may apply the multi-block ADMM extended directly from the ADMM from solving the 2-block convex optimization problem (Boyd et al., 2011). However, it has been shown in (Chen et al., 2013) that, different from the 2-block ADMM, the directly extended multi-block ADMM may diverge. There have been many existing works to deal with the possible non-convergence of multi-block ADMM. However, as pointed out in (Sun et al., 2014), for all existing various multi-block ADMM, although theoretically convergent, they usually performs substantially worse than the directly extended multi-block ADMM. To get a convergent multi-block ADMM that performs as efficiently as the directly extended ADMM, a proximal ADMM has been proposed in (Sun et al., 2014), which has both theoretical convergence guarantee and superior numerical efficiency over the directly extended ADMM.

For a given $\sigma > 0$, let

$$L_\sigma(Z, \Theta, G, W; \Lambda_1, \Lambda_2) := \|Z\|_{2,1} + \lambda \Omega_{\mathcal{G}'}(\Theta) + \rho \|G\|_1 + \langle \Lambda_1, Z + G + XW - Y \rangle \\ + \langle \Lambda_2, \Theta - CW \rangle + \frac{\sigma}{2} \|Z + G + XW - Y\|_F^2 + \frac{\sigma}{2} \|\Theta - CW\|_F^2 \tag{12}$$

be the augmented Lagrangian function for (10). Then apply the proximal ADMM proposed in (Sun et al., 2014) to optimization problem (10), giving rise to the following steps:

$$(Z^{k+1}, \Theta^{k+1}) = \arg\min L_\sigma(Z, \Theta, G^k, W^k; \Lambda_1^k, \Lambda_2^k), \tag{13}$$

$$W^{k+\frac{1}{2}} = \arg\min L_\sigma(Z^{k+1}, \Theta^{k+1}, G^k, W; \Lambda_1^k, \Lambda_2^k), \tag{14}$$

$$G^{k+1} = \arg\min L_\sigma(Z^{k+1}, \Theta^{k+1}, G, W^{k+\frac{1}{2}}; \Lambda_1^k, \Lambda_2^k), \tag{15}$$

$$W^{k+1} = \arg\min L_\sigma(Z^{k+1}, \Theta^{k+1}, G^{k+1}, W; \Lambda_1^k, \Lambda_2^k), \tag{16}$$

$$\Lambda_1^{k+1} = \Lambda_1^k + \tau\sigma(Z^{k+1} + XW^{k+1} + G^{k+1} - Y), \tag{17}$$

$$\Lambda_2^{k+1} = \Lambda_2^k + \tau\sigma(\Theta^{k+1} - CW^{k+1}), \tag{18}$$

where $\sigma > 0$, $\tau > 0$ are given parameters and the initial values of W, Λ_1 and Λ_2 should be selected such that $W^0 := (X^\top X + C^\top C)^{-1}(X^\top(Y - Z^0 - G^0) + C^\top \Theta^0)$ and $X^\top \Lambda_1^0 - C^\top \Lambda_2^0 = 0$, respectively.

Algorithm 1 Our *CMRG* algorithm

Input: Y, X, $\lambda > 0$, $\rho >$, $\sigma > 0$ and $\tau > 0$.
Initialization: Z^0, Θ^0, W^0, G^0, Λ_1^0, Λ_2^0 such that $W^0 := (X^\top X + C^\top C)^{-1}(X^\top(Y - Z^0 - G^0) + C^\top \Theta^0)$ and $X^\top \Lambda_1^0 - C^\top \Lambda_2^0 = 0$. $k = 0$.
repeat
  Compute Z^{k+1} using (19).
  Compute Θ^{k+1} using (20).
  Compute $W^{k+\frac{1}{2}}$ using (21).
  Compute G^{k+1} using (22).
  Compute W^{k+1} using (23).
  Compute Λ_1^{k+1} and Λ_2^{k+1} using (17).
  $k \leftarrow k + 1$.
until *Convergence*
Output: W^k, G^k.

All subproblems in (13)–(16) have closed form solutions. The solution Z^{k+1} to subproblem (13) is given by

$$(Z^{k+1})_{*j} = \left(1 - \frac{1}{\sigma\|(\Delta_Z^k)_{*j}\|_2}\right)_+ (\Delta_Z^k)_{*j}, \tag{19}$$

where $\Delta_Z^k := Y - XW^k - G^k - \frac{\Lambda_1^k}{\sigma}$. The solution Θ^{k+1} is given by

$$(\Theta^{k+1})_{g*} = \left(1 - \frac{\lambda}{\sigma\|(\Delta_\Theta^k)_{g*}\|_F}\right)_+ (\Delta_\Theta^k)_{g*},\ g \in \mathcal{G}' \tag{20}$$

where $\Delta_\Theta^k := CW^k - \frac{\Lambda_2^k}{\sigma}$. Let $A := X^\top X + D$, the solution $W^{k+\frac{1}{2}}$ to subproblem (14) is given by

$$W^{k+\frac{1}{2}} = A^{-1}(X^\top(Y - Z^{k+1} - G^k) + C^\top \Theta^{k+1}). \tag{21}$$

The solution G^{k+1} to subproblem (15) is given by

$$G_{ij}^{k+1} = \mathrm{Sign}((\Delta_G^k)_{ij}) \max\{(\Delta_G^k)_{ij} - \rho/\sigma, 0\}, \tag{22}$$

where $\Delta_G^k = Y - XW^{k+\frac{1}{2}} - Z^{k+1} - \frac{\Lambda_1^k}{\sigma}$ and $\mathrm{Sign}(\cdot)$ is the sign function. The solution W^{k+1} to subproblem (16) is given by

$$W^{k+1} = A^{-1}(X^\top(Y - Z^{k+1} - G^{k+1}) + C^\top \Theta^{k+1}). \tag{23}$$

Summarizing the above procedure leads to **Algorithm 1**. Compared with directly extended 3-block ADMM, the above proximal ADMM has an extra step for computing $W^{k+\frac{1}{2}}$. As seen from (21), the extra cost of computing $W^{k+\frac{1}{2}}$ is marginal provided that the Choleskey factorization of A is available. The reward for performing the extra step is the convergence of **Algorithm 1** as presented in **Theorem 1**. The proof of **Theorem 1** can be derived directly from Theorem 2.2 in (Sun et al., 2014).

Theorem 1 *Under the condition* $\tau \in \big(0, (1+\sqrt{5})/2\big)$*, the sequence* $\{(Z^k, \Theta^k, G^k, W^k, \Lambda_1^k, \Lambda_2^k)\}$ *generated by* ***Algorithm 1*** *converges to an unique point* $(\hat{Z}, \hat{\Theta}, \hat{G}, \hat{W}, \hat{\Lambda}_1, \hat{\Lambda}_2)$ *so that* $(\hat{Z}, \hat{\Theta}, \hat{G}, \hat{W})$ *solves optimization problem* (10) *and* $(\hat{\Lambda}_1, \hat{\Lambda}_2)$ *solves the dual problem of optimization problem* (10)*.*

4. Experiments

We compare our newly proposed model (8), which we denote as CMRG, with three existing models: the ordinary multivariate regression (OMR) model (5), the calibrated multivariate regression (CMR) model (7) and the ordinary multivariate regression with gross error (OMRG) model (9). We apply **Algorithm 1** to solve our model, and all other models are also solved similarly by ADMM.

4.1. Synthetic Data

To evaluate the finite-sample performance of our new model, we generate a simulation data following a similar scheme as in (Liu et al., 2013, 2014). Each data set consists of 400 training samples, 400 validation samples and 10,000 testing samples. Specifically,

1. Generate each row of X, independently from a 1000-dimensional normal distribution $N(0, \Sigma)$ where $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.5$ for all $i \neq j$.

2. Design a group sparsity structure so that $\mathcal{G} = \{\{1, \cdots, 100\}, \{91, \cdots, 190\}, \cdots, \{361, \cdots, 460\}, \{461\}, \cdots, \{1000\}\}$ and generate the regression coefficient matrix W so that entries in the first 460 rows are independently sampled from $N(0, 1)$ and all other entries are 0.

3. Generate the random noise matrix $Z = BD_0$, where $B \in \mathbb{R}^{1000\times 13}$ whose entries are independently sampled from $N(0, 1)$ and $D_0 \in \mathbb{R}^{13\times 13}$ is a diagonal matrix defined as $D_0 := \sigma_{max} \cdot \mathrm{diag}(2^{0/4}, 2^{-1/4}, \cdots, 2^{-12/4})$.

4. Generate the gross error matrix $G \in \mathbb{R}^{400\times 13}$ with S nonzero entries, whose positions are randomly selected and values are independently sampled from uniform distribution $\delta \cdot U(\sigma_{max}, 2\sigma_{max})$, where $\delta > 1$ is a scaling factor.

We conduct experiments with different values of σ_{max}, S and δ to evaluate the performance of CMRG. The parameter λ is chosen from $\lambda_0 * \{2^{-3}, 2^{-2.75}, \cdots, 2^{1.75}, 2^2\}$ with $\lambda_0 = \sqrt{\ln d} + \sqrt{p}$, and the parameter ρ is chosen from $\{10^{-3}, 10^{-2.8}, \cdots, 10^{0.8}, 10^1\}$. The optimal parameter $(\hat{\lambda}, \hat{\rho})$ is determined as

$$(\hat{\lambda}, \hat{\rho}) = \arg\min_{\lambda,\rho} \|\tilde{Y} - \tilde{X}\hat{W}^{\lambda,\rho}\|_F^2,$$

where $\hat{W}^{\lambda,\rho}$ is the resulted estimate using parameter (λ, ρ), and $\tilde{X}$ and $\tilde{Y}$ denote the design and response matrices of the validation data.

To evaluate the empirical performance, we adopt the following criteria:

$$\begin{aligned}
\text{Pre.Err.} &= \|\overline{Y} - \overline{X}\hat{W}\|_F / \|\overline{Y}\|_F, \\
\text{Adj.Pre.Err.} &= \|(\overline{Y} - \overline{X}\hat{W})D_0^{-1}\|_F / \|\overline{Y}D_0^{-1}\|_F, \\
\text{Est.Err.of W} &= \|W^* - \hat{W}\|_F / \|W^*\|_F, \\
\text{Est.Err.of G} &= \|G^* - \hat{G}\|_F / \|G^*\|_F,
\end{aligned}$$

where $(\overline{X}, \overline{Y})$ denotes the testing data, *Rec.Rate.of* W and *Rec.Rate.of* G measure the percentage of correctly recovered signs of entries in W^* and G^*, respectively.

Table 1 and Table 2 summarize the results averaged over 10 repetitions for different values of $(\sigma_{max}, S, \delta)$. We observe that regression models that take gross error into considerations (OMRG and CMRG) consistently outperform those that fail to do so (OMR and CMR), and regression models adopting the $\ell_{2,1}$-norm as the loss function (CMR and CMRG) perform better than those adopting least square loss (OMR and OMRG). Comparing CMRG with OMRG, we find that CMRG has lower prediction error and estimation error of W in most scenarios while OMRG has lower error rate in estimating G. Both CMRG and OMRG can successfully recover the sign of entries in W^* and G^* when only a small portion of observations are corrupted by gross errors, but the success rate decreases when the number of grossly corrupted observations is large. We believe that there exists some threshold so that when the ratio between sample size and the number of corrupted entries (i.e. n/S) is larger than the threshold, our approach can successfully recover the signed support of both W^* and G^*.

4.2. Big-Five Personality Prediction

The data set we use is a SNS data set built from microblogging site Sina Weibo (the Chinese counterpart of Twitter). To acquire the annotated data, we have developed a mental illness treatment website, *Anonymous*. By allowing online users to log in Weibo accounts through our *Anonymous* web platform, we are able to collect users' historical behavior data from Weibo into *Anonymous*. In total, there are 45 Weibo behavior features, which are categorized into four groups: *profile*, *self-presentation*, *security setting* and *social networking*, and 630 Weibo participants are recruited during the first two months of 2012.

We further scrutinize the participants to retain only those who are active users. We exclude users who either publish less than 512 blogs altogether, or have no blog published in the last three months. As a result, the data set contains 562 Weibo participants (instances), which is further split into disjoint training and testing sets, that contain 450 and 112 instances, respectively. To acquire ground-truth labels, each participant is additionally requested to fill up an inventory—the widely-used Big-Five Inventory from Berkeley Personality Lab. In practice, each element in this vector is a decimal number within the range of $[1, 5]$.

We compare our method with OMR, CMR, OMRG and ridge regression (RR) (i.e. model (2) with least square loss and $R(W) = \|W\|_F^2$). Similar to synthetic data, we choose parameter λ from $\lambda_0 * \{10^{-3}, 10^{-2.5}, \cdots, 10^2\}$ with $\lambda_0 = \sqrt{\ln d} + \sqrt{p}$, and the parameter ρ from $\{10^{-2}, 10^{-1.5}, \cdots, 10^3\}$. The optimal parameter $(\hat{\lambda}, \hat{\rho})$ is determined by 5-fold cross-

Table 1: Statistical performance (mean±std) of four regression models: OMR, CMR, OMRG and CMRG.

		Method	*Pre.Err.*	*Adj.Pre.Err.*	*Est.Err.of W*	*Rec.Rate.of W*	*Est.Err.of G*	*Rec.Rate.of G*
$\sigma_{max} = \sqrt{2}$								
$S = 50$	$\delta = 5$	OMR	0.3089±0.0020	0.2930±0.0025	0.4123±0.0022	93.81±0.12	-	-
		CMR	0.3086±0.0015	0.2931±0.0011	0.4115±0.0014	93.84±0.12	-	-
		OMRG	**0.3075**±0.0016	**0.2914**±0.0009	**0.4104**±0.0015	93.86±0.03	**0.9986**±0.0279	**98.18**±0.68
		CMRG	0.3082±0.0017	0.2920±0.0012	0.4112±0.0015	**93.87**±0.09	1.0104±0.0398	98.08±1.62
	$\delta = 10$	OMR	0.3636±0.0076	0.3381±0.0015	0.4856±0.0109	92.85±0.47	-	-
		CMR	0.3454±0.0040	0.3251±0.0017	0.4605±0.0054	93.16±0.24	-	-
		OMRG	0.3244±0.0007	0.3045±0.0024	0.4328±0.0006	93.51±0.08	**0.7177**±0.0469	**95.96**±0.00
		CMRG	**0.3243**±0.0009	**0.3034**±0.0028	**0.4328**±0.0007	**93.59**±0.17	0.7237±0.0468	94.65±0.16
$S = 500$	$\delta = 5$	OMR	0.4635±0.0060	0.4422±0.0103	0.6193±0.0085	91.07±0.20	-	-
		CMR	0.3930±0.0025	0.3739±0.0037	0.5241±0.0031	92.11±0.10	-	-
		OMRG	0.3968±0.0019	0.3784±0.0041	0.5292±0.0027	92.02±0.18	**0.9543**±0.0085	85.58±0.74
		CMRG	**0.3918**±0.0020	**0.3730**±0.0044	**0.5224**±0.0027	**92.13**±0.11	0.9552±0.0205	**89.51**±0.60
	$\delta = 10$	OMR	0.8054±0.0094	0.7898±0.0060	1.0756±0.0132	86.96±0.01	-	-
		CMR	0.4883±0.0046	0.4725±0.0001	0.6501±0.0062	89.99±0.08	-	-
		OMRG	0.4636±0.0022	0.4498±0.0035	**0.6176**±0.0026	85.74±1.31	0.7189±0.0242	**82.70**±1.02
		CMRG	**0.4635**±0.0020	**0.4496**±0.0038	**0.6176**±0.0023	**90.48**±0.09	**0.7187**±0.0247	78.89±0.75
$S = 2500$	$\delta = 5$	OMR	0.8608±0.0092	0.8224±0.0196	1.1477±0.0135	85.59±0.26	-	-
		CMR	**0.5075**±0.0013	**0.4862**±0.0024	**0.6755**±0.0021	**89.36**±0.09	-	-
		OMRG	0.5346±0.0020	0.5138±0.0047	0.7117±0.0030	45.78±14.55	**0.9274**±0.0092	**60.20**±0.46
		CMRG	**0.5075**±0.0013	**0.4862**±0.0024	**0.6755**±0.0021	**89.36**±0.09	1.0000±0.0000	51.92±0.00
	$\delta = 10$	OMR	1.4958±0.0196	1.4361±0.0516	2.0030±0.0248	76.81±0.87	-	-
		CMR	**0.6120**±0.0004	**0.5876**±0.0023	**0.8126**±8.3e-5	**86.23**±0.13	-	-
		OMRG	0.6391±0.0019	0.6139±0.0047	0.8465±0.0007	40.22±2.49	**0.6965**±0.0019	**65.89**±0.07
		CMRG	**0.6120**±0.0004	**0.5876**±0.0023	**0.8126**±8.3e-5	**86.23**±0.13	1.0000±0.0000	51.92±0.00

Table 2: Statistical performance (mean±std) of four regression models: OMR, CMR, OMRG and CMRG.

		Method	*Pre.Err.*	*Adj.Pre.Err.*	*Est.Err.of W*	*Rec.Rate.of W*	*Est.Err.of G*	*Rec.Rate.of G*
$\sigma_{max} = 2\sqrt{2}$								
$S = 50$	$\delta = 5$	OMR	0.3715±0.0024	0.3459±0.0058	0.4958±0.0017	92.73±0.09	-	-
		CMR	0.3596±0.0008	0.3380±0.0042	0.4800±2.7e-5	92.87±0.02	-	-
		OMRG	0.3451±0.0008	0.3191±0.0072	0.4608±0.0023	**93.19**±0.01	0.7867±0.0179	**95.20**±0.20
		CMRG	**0.3439**±0.0011	**0.3179**±0.0071	**0.4593**±0.0027	93.16±0.13	**0.7850**±0.0201	94.74±0.15
	$\delta = 10$	OMR	0.5100±0.0029	0.4877±0.0207	0.6801±0.0057	90.51±0.02	-	-
		CMR	0.4198±0.0001	0.4021±0.0087	0.5593±0.0012	91.55±0.29	-	-
		OMRG	0.3471±0.0028	0.3223±0.0019	0.4632±0.0028	**93.12**±0.02	**0.4205**±0.0022	**95.04**±0.44
		CMRG	**0.3470**±0.0023	**0.3204**±0.0005	**0.4630**±0.0019	**93.12**±0.08	0.4245±0.0096	93.53±2.35
$S = 500$	$\delta = 5$	OMR	0.7308±0.0054	0.6789±0.0035	0.9772±0.0079	87.48±0.16	-	-
		CMR	0.4890±0.0016	0.4661±0.0031	0.6513±0.0025	89.86±0.18	-	-
		OMRG	**0.4701**±0.0022	**0.4513**±0.0028	**0.6266**±0.0030	**90.32**±0.24	**0.7554**±0.0023	**79.58**±0.14
		CMRG	0.4708±0.0022	0.4521±0.0024	0.6278±0.0032	**90.32**±0.32	0.7570±0.0018	77.67±0.05
	$\delta = 10$	OMR	1.4450±0.0049	1.4221±0.0092	1.9346±0.0072	82.44±0.33	-	-
		CMR	0.5967±0.0005	0.5772±0.0040	0.7953±0.0003	86.78±0.10	-	-
		OMRG	0.4920±0.0029	0.4707±0.0019	0.6557±0.0048	78.71±15.26	0.4208±0.0097	**77.05**±2.57
		CMRG	**0.4918**±0.0031	**0.4699**±0.0016	**0.6554**±0.0049	**89.69**±0.20	**0.4203**±0.0094	75.71±0.76
$S = 2500$	$\delta = 5$	OMR	1.5402±0.0176	1.4759±0.0016	2.0561±0.0228	81.31±0.59	-	-
		CMR	**0.6118**±0.0006	**0.5902**±0.0024	**0.8145**±0.0018	**86.36**±0.29	-	-
		OMRG	0.6388±0.0011	0.6177±0.0077	0.8503±0.0035	39.42±1.13	**0.7049**±0.0090	**66.37**±0.14
		CMRG	**0.6118**±0.0006	**0.5902**±0.0024	**0.8145**±0.0018	**86.36**±0.30	1.0000±0.0000	51.92±0.00
	$\delta = 10$	OMR	3.0067±0.1150	2.9223±0.1457	4.0145±0.1506	73.14±0.26	-	-
		CMR	0.7021±2.8e-6	0.6786±0.0015	0.9295±0.0026	82.92±0.29	-	-
		OMRG	0.6859±0.0017	0.6615±0.0025	0.9093±0.0024	83.23±0.42	0.4126±0.0019	**66.35**±1.99
		CMRG	**0.6856**±0.0015	**0.6607**±0.0017	**0.9085**±0.0020	**83.35**±0.30	**0.4124**±0.0020	65.62±0.41

Table 3: Prediction error on Weibo data without gross error.

	Weibo data without gross error					Corrupted Weibo data				
	RR	OMR	CMR	OMRG	CMRG	RR	OMR	CMR	OMRG	CMRG
Agr.	0.1784	0.1788	**0.1783**	0.1788	**0.1783**	0.2176	0.2146	0.2136	0.2055	**0.1914**
Con.	**0.2128**	0.2226	0.2212	0.2226	0.2212	0.2332	0.2174	0.2170	0.2160	**0.2109**
Ext.	**0.2147**	0.2172	0.2152	0.2172	0.2152	0.2384	0.2340	0.2379	0.2310	**0.2205**
Neu.	**0.2262**	0.2269	0.2271	0.2269	0.2271	0.2670	0.2676	0.2622	0.2594	**0.2543**
Ope.	**0.1717**	0.1830	0.1823	0.1830	0.1823	0.2088	0.1822	0.1814	0.1720	**0.1641**
Pre.Err.	**0.1993**	0.2046	0.2037	0.2046	0.2037	0.2320	0.2231	0.2221	0.2166	**0.2076**

Table 4: Recovery accuracy of OMRG and CMRG.

Method	$Est.Err.of\ G$	$Rec.Rate.of\ G$
OMRG	0.8859±0.2552	89.23±0.76
CMRG	**0.7130**±0.0052	**99.20**±0.30

validation on the training data. For performance evaluation, we use the relative prediction error as in synthetic data.

We first conduct experiments on the training data without gross error. Averaged results over 10 repetitions are shown in Table 3, where abbreviations in the left-most column denote *Agreeableness*, *Conscientiousness*, *Extraversion*, *Neuroticism* and *Openness*, respectively, and *Pre.Err.* denotes the relative prediction error for all five personalities. We observe from Table 3 that regression models considering gross error (OMRG and CMRG) obtain the same results as those that fail to do so (OMR and CMR), since there is no gross error in the training data. Moreover, CMRG has better performance than regression models OMR and OMRG while it is slightly worse than ridge regression.

To model practical scenarios where some observations may be missing, we randomly delete 250 entries ($\approx 10\%$) in the responses and conduct experiments on the corrupted training data. Average prediction errors over 10 repetitions are shown in Table 3 and Table 4 that shows the accuracy of OMRG and CMRG for recovering the deleted entries. We can see that CMRG performs significantly better than other methods. More importantly, CMRG is capable of recovering missing observations accurately as shown in Table 4.

As a reference, we also compute the averaged absolute distance (AAD) between ground-truth and real-valued prediction based on the corrupted data for all regression models, and compare them with the best AAD results of (Golbeck et al., 2011) on Twitter dataset. Averaged results over 10 repetitions are shown in Table 5, where the last column lists the results of (Golbeck et al., 2011). We see that CMRG is the best and much better than the best AAD results of (Golbeck et al., 2011) on Twitter dataset.

5. Conclusions

A new approach has been proposed to address the problem of multivariate regression with missing and grossly corrupted observations. Our approach takes gross error into consideration and at the same time calibrates each regression task with respect to its noise level. An efficient and convergent proximal ADMM method has also been proposed to solve the

Table 5: Comparison of AAD results on the corrupted data.

	RR	OMR	CMR	OMRG	CMRG	(Golbeck et al., 2011)
Agr.	0.65	0.64	0.64	0.61	**0.56**	0.65
Con.	0.59	0.55	0.55	0.55	**0.54**	0.73
Ext.	0.62	0.62	0.63	0.61	**0.59**	0.80
Neu.	0.68	0.68	0.66	0.65	**0.64**	0.91
Ope.	0.61	0.53	0.53	0.50	**0.48**	0.60
Ave.	0.63	0.60	0.60	0.58	**0.56**	0.74

induced optimization problem. Moreover, our method can be successfully applied to predict personalities based on behaviors at SNSs. Experiments on synthetic and real data corroborate the effectiveness of our algorithm. In future, we plan to investigate statistical properties of our method and the effect of different behaviors at SNSs to the prediction of personality.

Acknowledgments

This research was supported by A*STAR JCO grants 1231BFG040 and 1431AFG120.

References

T. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73:243–272, 2008.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Optimization Online*, 2013.

Xi Chen, Qihang Lin, Seyoung Kim, Jaime G. Carbonell, and Eric P. Xing. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6:719–752, 2012.

T. Correa, A. Hinsley, and H. Ziga. Who interacts on the web? the intersection of users' personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2010.

G. Domino and M. Domino. *Psychological testing: An introduction*. Cambridge University Press, 2006.

Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

D. Funder. Personality. *Annu. Rev. Psychol.*, 52:197–221, 2001.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *IEEE International Conference on Social Computing*, 2011.

S. Gosling, A. Augustine, S. Vazire, N. Holtzman, , and S. Gaddis. Manifestations of personality in online social networks: Self-reported facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking*, 14:483–488, 2011.

X. Jie. Sequencing algorithm based on the social network real-time search engin. *Science Technology and Engineering*, 28:1671–1815, 2011.

R. Landers and J. Lounsbury. An investigation of big-five and narrow personality traits in relation to internet usage. *Computers in Human Behavior*, 22(2):283–293, 2006.

X. Li. compressed sensing and matrix completion with constant proportion of corruptions. Technical report, arXiv:1104.1041v2, 2012.

H. Liu, L. Wang, and T. Zhao. Multivariate regression with calibration. Technical report, arXiv:1305.2238v1, 2013.

H. Liu, L. Wang, and T. Zhao. Multivariate regression with calibration. In *NIPS*, 2014.

Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39: 2164–2204, 2011.

S. Ma, C. Jiao, , and M. Zhang. Application of social network analysis in psychology. *Advances in Psychological Science*, 19(5):755–764, 2011.

G. Matthews, I. Deary, and M. Whiteman. *Personality Traits*. Cambridge University Press, 2006.

Nam H. Nguyen and Trac D. Tran. Robust lasso with missing and grossly corrupted observations. *IEEE Transactions on Information Theory*, 59:2036–2058, 2013.

E. Orr, M. Sisic, C. Ross, M. Simmering, J. Arseneault, and R. Orr. The influence of shyness on the use of facebook in an undergraduate sample. *Cyberpsychology and Behavior*, 12: 337–340, 2009.

Z. Qin and D. Goldfarb. Structured sparsity via alternating direction methods. *Journal of Machine Learning Research*, 13:1435–1468, 2012.

J. Reynol. The relationship between frequency of facebook use, participation in facebook activities, and student engagement. *Computers and Education*, 58:162–171, 2011.

A. Rohde and A. B. Tsybakov. Estimation of hign-dimensional low-rank matrices. *The Annals of Statistics*, 39:887–930, 2011.

Defeng Sun, Kim-Chuan Toh, and Liuqin Yang. A convergent proximal alternating direction method of multipliers for conic programming with 4-block constraints. Technical report, arXiv:1404.5378, 2014.

John Wright and Yi Ma. Dense error correction via l1-minimization. *IEEE Transactions on Information Theory*, 56:3540–3560, 2010.

Huan Xu and Chenlei Leng. Robust multi-task regression with grossly corrupted observations. In *AISTATS*, 2012.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 58:3047–3064, 2012.

Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust pca: The high-dimensional case. *IEEE Transactions on Information Theory*, 59:546–572, 2013.