Robust Multi-Factor Personality Prediction with Correlated Data Corruption in Social Media

ABSTRACT

Personality prediction in multiple factors, such as openness and agreeableness, is growing in interest especially in the context of social media, which contains massive online posts or likes that can potentially reveal an individual's personality. However, the data collected from social media inevitably contains massive amounts of noise and corruption. To address it, traditional robust methods still suffer from several important challenges, including 1) existence of correlated corruption among multiple factors, 2) difficulty in estimating the corruption ratio in multi-factor data, and 3) scalability to massive datasets. This paper proposes a novel robust multi-factor personality prediction model that concurrently addresses all the above challenges by developing a distributed robust regression algorithm. Specifically, the algorithm optimizes regression coefficients of each factor in parallel with a heuristically estimated corruption ratio and then consolidates the uncorrupted set from multiple factors in two strategies: global consensus and majority voting. We also prove that our algorithm benefits from strong guarantees in terms of convergence rates and coefficient recovery, which can be utilized as a generic framework for the multi-factor robust regression problem with correlated corruption property. Extensive experiment on synthetic and real dataset demonstrates that our algorithm is superior to those of existing methods in both effectiveness and efficiency.

1 INTRODUCTION

Personality is the particular combination of emotional, attitudinal, and behavioral response patterns accounting for individual differences in people [1]. The classic Five-Factor model [2] comprises five traits: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. Individual personality analysis is important and widely used in many real-world applications with characterized services [3]. For instance, an *extraversive* user may have a higher frequency of online activities and be more likely to use a recommendation system to make new friends with strangers. To identify an individual's personality, the most commonly used method is self-report inventory [4], which requires individuals to answer questions about their typical behavior. For example, the Berkeley Personality Lab has designed a widely used Big-Five Inventory (BFI) containing 44 questions to form reliable five-factor personality scores. Despite its wide usage, the self-report method still suffers from two major problems: It is particularly difficult to conduct in large scale, and *social desirability* [5] may influence responses in that people might state what they wish, which will reduce the credibility of their responses.

Recently, the growing popularity of social media has provided a new way to manifest personality through users' online activities, which yields important insights into users' interests, preferences, and sentiments. As online behaviors share a significant amount in common with real-world behaviors [6], social media provides an excellent data source to reproduce social activities in real life from a large and diverse population. Although social media can easily extend the traditional methods to large scale, the personality data collected from social media is more easily corrupted by *careless response* [7] and *social desirability* [5] due to the lack of supervision. For instance, one may distort the response of personality assessment to exaggerate the personality score of *conscientiousness* due to one's social desirability; or randomly respond the questions in personality assessment carelessly.

Existing work on personality prediction in social media typically only focuses on building the relation between personality and online behaviors, although careless or malicious user annotations are actually a crucial issue in practice [5]. For those seeking to address this issue, the major challenges can be summarized as follows. 1) **Existence of correlated corruption**. Featured by the characteristics of personality, factors of personality have corruption correlations. For example, when one tends to fake a higher score of *extraversion* in one personality test, the *neuroticism* score will be correspondingly impacted [8]. Thus, simply estimating corruption independently for each factor is not an ideal strategy, as their interactions also need to be considered. 2) **Difficulty in estimating the corruption ratio**. Existing methods typically assume the corruption ratio is a user known parameter; however, the parameter can hardly be estimated in practice even when users obtain the corresponding domain knowledge. Moreover, due to the existence of correlated corruption among the factors, as mentioned earlier, corruption ratio are not independent of each other. It is thus clearly necessary to utilize this correlation pattern to regulate the corruption estimation process. 3) **Scalability to massive datasets**. Different from the traditional inventory-based psychological analysis based on at most hundreds of data samples, millions of samples are being generated by social media users everyday. Therefore, considering the complexity of robust personality prediction problem, an extremely efficient algorithm is required to handle the massive datasets.

In order to simultaneously address all these technical challenges, this paper presents a novel model based on multi-factor learning, **Robust Multi-Factor Personality Prediction (RMFP)**, which jointly optimizes the regression coefficients with correlated corruption. In our RMFP, the uncorrupted sample set for the prediction of each factor will be consolidated into a unified set. This improves the estimation of not only the regression coefficients but also the corruption patterns. Moreover, by letting the factors learn the estimation of corruption from each other, our algorithm achieves faster convergence than optimizing them individually. In addition, each factor can be optimized in parallel, which is extremely beneficial to efficiency when the number of factors becomes large. The main contributions of this paper are as follows:

- **Formulating a model for robust multi-factor personality prediction.** The proposed model considers correlation of multiple personality factors by utilizing the correlated corruption property. Based on this property, each personality factor learns the estimated corruption pattern from the others to improve overall performance.
- **Proposing a distributed robust algorithm for the multi-factor regression problem.** The optimization of the proposed multi-factor model is a non-convex discrete optimization problem, which is technically challenging. By optimizing

individual factor in parallel, the uncorrupted set is combined from each factor following two strategies: global consensus and majority voting.

- **Providing a strong recovery guarantee under the multi-factor problem setting.** We prove that our RMFP algorithm with global consensus strategy converges at a geometric rate and recovers coefficients of each factor exactly under the assumption of Subset Strong Convexity and Subset Strong Smoothness properties. Specifically, we prove that our algorithm ensures a close error bound of the regression coefficient compared to ground truth.
- **Conducting extensive experiments for performance evaluations.** The proposed method was evaluated on both synthetic data and real-world datasets with various corruption settings. The results demonstrate that the proposed approach runs efficiently and consistently outperforms the best of the existing methods along multiple metrics.

The rest of this paper is organized as follows. Section 2 reviews background and related work, and Section 3 introduces the problem setup. The proposed RMFP algorithm is presented in Section 4. Section 5 presents the proof of convergence rate and recovery guarantee. The experiments on both synthetic and real-world datasets are presented in Section 6, and the paper concludes with a summary of the research in Section 7.

2 RELATED WORK

The work related to this paper falls into three categories and is summarized below.

Personality prediction in social media: Many approaches [9][10] [11] that combine personality and social media have been proposed. Some research focuses on the association relation between personality and behavior rather than quantitative analysis of personality. For instance, Orr et al. [9] discovered the relation between shyness and the number of "friends", while Correa et al. [12] found extraversion and openness have a positive relation to the user experience in social media. A large body of work explores linguistic feature selection in personality prediction. The studies focus on a broad set of demographic and psychological features in user postings such as social status [10][13], occupation [14], mental illness [15][16], political orientation [17], and gender [18], as well as other online behavior features such as Facebook likes [19] and profile pictures [20][21]. However, none of these works considers the noise or corruption in the social media dataset. Recently, Zhang et al. [11] proposed the only work that considers the robustness of the personality prediction model. However, the model treats the corruption of different factors independently and does not consider the correlation between the corruption of multiple factors.

Robust regression model: A large body of literature on the robust regression problem has been built up over the last few decades. Most of them lack the theoretical guarantee of regression coefficients recovery [22][23][24]. To pursue the exact recovery results for the robust regression problem, some work focused on L_1 penalty-based convex formulations [25][26]. However, these methods imposed severe restrictions on the data distribution such as row-sampling from an incoherent orthogonal matrix[26]. Several studies require the corruption ratio parameter, which is difficult to determine manually. For instance, She and Owen [27] rely on a regularization parameter to control the size of the uncorrupted set based on soft-thresholding. Instead of a regularization parameter, Chen et al. [28] require the upper bound of outliers number, which is also difficult to estimate. Recently, Bhatia et al. [29] proposed a hard-thresholding algorithm. Although the method gives a strong guarantee of coefficient recovery under a mild assumption on input data, their results are highly dependent on the corruption ratio parameter inputed by users. Zhang et al. [11] consider the robust multivariate regression problem in personality prediction with gross corruption; however, their method also requires regularization parameters to control the ratio and sparsity of corruption. None of these approaches focuses specifically on the robust multi-factor regression problem with a correlated corruption ratio under a strong recovery guarantee.

Multi-task learning: Multiple related tasks are simultaneously learned in multi-task learning (MTL) to prove the generalization performance [30]. Many multi-task approaches [31][32][33] have been proposed in recent years to model the relationship between tasks. The relatedness of tasks can be characterized by constraining multiple tasks to share a common underlying structure, such as a common subspace [31][33], a common set of features [34], or using a structured model [35]. For example, Evgeniou et al. [33] proposed a regularized multi-task learning method that constrained the models of all the tasks to be close to each other. MTL approaches have been applied in many domains, including computer vision and biomedical informatics. To the best of our knowledge, however, ours is the first work that applies MTL in the domain of robust personality prediction.

3 PROBLEM SETTING

In this section, the problem addressed by this research is formulated.

Denote $X = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]$ as a collection of social media features for N users, where each column $\boldsymbol{x}_i \in \mathbb{R}^{P\times 1}$ represents the feature set for the i^{th} user, and $\boldsymbol{y} = \{\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, ..., \boldsymbol{y}^{(M)}\}$ represents the personality labels for M factors, where $\boldsymbol{y}^{(m)} \in \mathbb{R}^{N\times 1}$ is the personality label for all the N users in the m^{th} factor. We assume the response vector $\boldsymbol{y}^{(m)}$ of the m^{th} factor is generated using the model:

$$\boldsymbol{y}^{(m)} = \boldsymbol{X}^T \boldsymbol{\beta}_*^{(m)} + \boldsymbol{u}^{(m)} + \boldsymbol{\varepsilon}^{(m)} \tag{1}$$

where $\boldsymbol{\beta}_*^{(m)}$ represents the ground truth coefficients of the regression model and $\boldsymbol{u}^{(m)} \in \mathbb{R}^{N\times 1}$ is the unbounded corruption vector introduced adversarily or unintentionally. When the i^{th} sample is uncorrupted, we have $\boldsymbol{u}_i^{(m)} = 0$; otherwise, $\boldsymbol{u}_i^{(m)}$ represents the corresponding corruption value. Denoting S_* as the set of uncorrupted samples in m^{th} factor, then we have $S_* = \overline{supp(\boldsymbol{u}^{(m)})}$. $\boldsymbol{\varepsilon}^{(m)}$ is denoted as the additive dense noise for the m^{th} factor, where $\varepsilon_i^{(m)} \sim \mathcal{N}(0, \sigma^2)$. The notations used in this paper are summarized in Table 1.

Before formally stating the problem, we first introduce two definitions related to corruption in personality assessment.

DEFINITION 1. ***Corruption in personality assessment.*** *Corruption or faking in personality assessment represents the response distortion aimed at providing a self-description that helps to achieve personal goals (e.g., social desirability [5], careless response [7]).*

Existing psychological research [8][36] shows that corruption indeed increases correlations between the Big Five factors in personality. The empirical evidence of these works concludes that faking on personality assessment can be considered a source of systematic variance that is added to (or even replaces) the true score variance. For example, a faked score on one particular test of *extraversion* has higher correlations with other factors of personality such as *conscientiousness* or *neuroticism* than scores on other *extraversion*

tests. Based on this observation, we formally define the correlated corruption property as follows.

DEFINITION 2. ***Correlated Corruption Property.*** *Denoting $\boldsymbol{u}^{(m)}$ as the corruption vector of the m^{th} factor, the M factors satisfy the correlated corruption property if the following holds:*

$$supp(\boldsymbol{u}^{(i)}) = supp(\boldsymbol{u}^{(j)}) \quad \forall i, j \in \{1...M\} \tag{2}$$

where $supp(\boldsymbol{u}^{(i)})$ denotes the set of corrupted points in the i^{th} factor.

Table 1: Math Notations

Notations	Explanations
$X \in \mathbb{R}^{F \times N}$	collection of social media feature set
$\boldsymbol{y}^{(m)} \in \mathbb{R}^{N \times 1}$	response vector of the m^{th} factor
$\boldsymbol{\beta}^{(m)} \in \mathbb{R}^{F \times 1}$	estimated regression coefficient of the m^{th} factor
$\boldsymbol{u}^{(m)} \in \mathbb{R}^{N \times 1}$	corruption vector of the m^{th} factor
$\boldsymbol{r}^{(m)} \in \mathbb{R}^{N \times 1}$	residual vector of the m^{th} factor
$S \subseteq [n]$	estimated uncorrupted set
$S_* \subseteq [n]$	ground truth uncorrupted set, where $S_* = \overline{supp(\boldsymbol{u})}$

The goal of our study is to learn a new multi-factor robust regression problem with correlated corruption, which jointly recovers the multi-factor regression coefficients $\hat{\boldsymbol{\beta}} = \{\hat{\boldsymbol{\beta}}^{(1)}...\hat{\boldsymbol{\beta}}^{(M)}\}$ and determines the correlated uncorrupted set $\hat{S}$ simultaneously. The problem is formally defined as follows:

$$\begin{aligned} &\hat{\boldsymbol{\beta}}, \hat{S} = \arg\min_{\boldsymbol{\beta}, S} \sum_{m=1}^{M} \|\boldsymbol{y}_S^{(m)} - X_S^T \boldsymbol{\beta}^{(m)}\|_2^2 \\ &s.t. \quad S \in \left\{\Gamma(Z) \,\middle|\, \forall m \in \{1, \dots, M\} : S^{(m)} \geq h(\boldsymbol{r}^{(m)})\right\} \end{aligned} \tag{3}$$

Denoting $S^{(m)}$ as the estimated uncorrupted set for the m^{th} factor and $Z = \{S^{(1)}, \dots, S^{(M)}\}$ as the collection of uncorrupted set for all the factors, the function $\Gamma(\cdot)$ consolidates the estimation of all the factors into one correlated uncorrupted set $S \subset [n]$. $\boldsymbol{y}_S$ restricts the row of $\boldsymbol{y}$ to indices in S, and X_S signifies that the columns of X are restricted to indices in S. Therefore, we have $\boldsymbol{y}_S^{(m)} \in \mathbb{R}^{|S| \times 1}$ and $X_S \in \mathbb{R}^{p \times |S|}$. The size of $S^{(m)}$ is lower-bounded by a heuristic function $h(\cdot)$ according to the m^{th} factor's residual vector $\boldsymbol{r}^{(m)} = \boldsymbol{y}^{(m)} - X^T \boldsymbol{\beta}^{(m)}$. Also, we use $\boldsymbol{r}_S^{(m)}$ to represent the $|S|$-dimensional residual vector containing the components in S for the m^{th} factor. The detail of heuristic function $h(\cdot)$ and consolidation function $\Gamma(\cdot)$ will be explained in Section 4.1 and 4.2.

4 RMFP MODEL

In this section, we propose a new model, Robust Multi-Factor Personality Prediction (RMFP), which is based on robust multi-factor regression. In Section 4.1, the corruption estimation for each factor is characterized mathematically by a heuristic hard thresholding method, and then these estimations are consolidated for multiple factors in Section 4.2. In Section 4.3, a novel distributed robust regression algorithm is proposed that ensures a strong guarantee of coefficient recovery.

4.1 Corruption Estimation

The size of uncorrupted set $S^{(m)}$ for m^{th} factor in Equation (3) is lower-bounded by heuristic function $h(\cdot)$, which is defined as follows.

$$h(\boldsymbol{r}^{(m)}) := \arg\max_{\tau \in \mathbb{Z}^+, \tau \leq n} \tau \quad s.t. \;\; r_{\delta(\tau)}^{(m)} \leq \frac{2\tau r_{\delta(\tau_o)}^{(m)}}{\tau_o} \tag{4}$$

The variable τ_o in the constraint is defined as follows:

$$\tau_o = \arg\min_{1 \leq \tau \leq n} \left| \left(r_{\delta(\tau)}^{(m)}\right)^2 - \frac{\|\boldsymbol{r}_{S_{\tau'}}^{(m)}\|_2^2}{\tau'} \right| \tag{5}$$

where $\tau' = \tau - \lceil n/2 \rceil$ and $S_{\tau'}$ is the position set containing the smallest τ' elements in residual $\boldsymbol{r}^{(m)}$, and $r_{\delta(k)}^{(m)}$ represents the k^{th} elements of $\boldsymbol{r}^{(m)}$ in ascending order of magnitude.

Basically, the design follows a natural intuition that data points with unbounded corruption always have a higher residual $r_i^{(m)} = y_i^{(m)} - X_i \boldsymbol{\beta}^{(m)}$ in magnitude compared to uncorrupted data. Moreover, the constraint in Equation (4) ensures the residual of the largest element τ in our estimation cannot be too much larger than the residual of a smaller element τ_o. This is because if the element τ_o is too small, some uncorrupted elements will be excluded from our estimation; otherwise, if it is too large, some corrupted elements will be included. The formal definition of τ_o is shown in Equation (5), in which τ_o is defined as a value whose squared residual is closest to $\|\boldsymbol{r}_{S_{\tau'}}^{(m)}\|_2^2/\tau'$, where τ' is less than the ground truth threshold τ_* as the corruption ratio is typically assumed not larger than half [28][29]. This design ensures that $|S_*^{(m)} \cap S_t^{(m)}| \geq \tau - n/2$, which means at least $\tau - n/2$ elements are correctly estimated in $S_t^{(m)}$. The property will be used in the proof in Lemma 1 in Section 5. In addition, the precision of the estimated uncorrupted set can be easily achieved when fewer elements are included in the estimation, but with low recall value. To increase the recall of our estimation, the objective function in Equation (4) chooses the maximum uncorrupted set size.

Applying the uncorrupted set size generated by $\tau(\cdot)$, the heuristic hard thresholding is defined as follows:

DEFINITION 3. ***Heuristic Hard Thresholding***. *Denoting $\delta_{\boldsymbol{r}}^{-1}(i)$ as the position of the i^{th} element in residual vector $\boldsymbol{r}$'s ascending order of magnitude, the heuristic hard thresholding of $\boldsymbol{r}$ is defined as*

$$\mathcal{H}_\tau(\boldsymbol{r}) = \{i \in [n] : \delta_{\boldsymbol{r}}^{-1}(i) \leq h(\boldsymbol{r})\} \tag{6}$$

The optimization of $S^{(m)}$ is formulated as solving Equation (6), where the set returned by $\mathcal{H}_\tau(\boldsymbol{r}^{(m)})$ will be used as the estimated uncorrupted set of the m^{th} factor.

4.2 Corruption Consolidation

To consolidate the uncorrupted sets of all the factors, we propose two strategies for the consolidation function $\Gamma(\cdot)$ in Equation (3): *global consensus* (*GC*) and *majority voting* (*MV*). The definition of global consensus strategy is presented as follows.

DEFINITION 4. ***Global Consensus.*** *Denoting $S^{(m)}$ as the uncorrupted set of the m^{th} factor, the global consensus uncorrupted set S is defined as $S = \bigcap_{m=1}^{M} S^{(m)}$.*

Global consensus strategy selects the estimated uncorrupted set as the intersection set from all the factors. The strategy strictly follows

Algorithm 1: RMFP ALGORITHM

Input: Multi-factor training dataset X, $\boldsymbol{y}$, tolerance ϵ
Output: solution $\hat{\boldsymbol{\beta}}$, $\hat{S}$

1 $S_0 \leftarrow [n]$, $\Psi \leftarrow [M]$, $t \leftarrow 0$
2 **repeat**
3 **for** $m \in \Psi$ **do**
4 $\boldsymbol{\beta}_{t+1}^{(m)} \leftarrow (X_{St} X_{St}^T)^{-1} X_{St} \boldsymbol{y}_{St}^{(m)}$
5 $\boldsymbol{r}_{t+1}^{(m)} \leftarrow |\boldsymbol{y}^{(m)} - X^T \boldsymbol{\beta}^{(m)}|$
6 $S_{t+1}^{(m)} \leftarrow \mathcal{H}_\tau(\boldsymbol{r}_{t+1}^{(m)})$ *// Heuristic hard threshold on residual*
7 **if** $\|\boldsymbol{r}_{t+1}^{(m)} - \boldsymbol{r}_t^{(m)}\|_2 / n < \epsilon$ **then**
8 $\Psi \leftarrow \Psi \setminus \{m\}$ *// Remove factor from active set*
9 $S_{t+1} = \Gamma(S_{t+1}^{(1)}, \ldots, S_{t+1}^{(M)})$ *// Corruption Consolidation: GC or MV*
10 $t \leftarrow t + 1$
11 **until** $\Psi = \emptyset$
12 **return** $\boldsymbol{\beta}_t^{(1)} \ldots \boldsymbol{\beta}_t^{(M)}$, S_t

the *correlated corruption property* (Definition 2) and only keeps the uncorrupted elements contained in estimators of all the factors. It increases the precision of the estimated uncorrupted set, but also excludes some correct samples from the estimation. As dense noise will increase the difference between the estimations of multiple factors, the strategy is more suitable to cases in which the amount of dense noise is small or there is no dense noise.

To avoid the problem of global consensus strategy, we propose the second strategy, majority voting, as follows.

DEFINITION 5. ***Majority Voting.*** *Denoting $S^{(m)}$ as the uncorrupted set of the m^{th} factor and M as the total number of factors, the uncorrupted set S of majority voting satisfies:*

$$\sum_{m=1}^{M} \mathbb{1}\left(S_i \in S^{(m)}\right) \geq \left\lceil \frac{M}{2} \right\rceil \quad \forall i \leq |S| \tag{7}$$

Different from global consensus, which requires the consolidated elements to be contained in estimators of all factors, the majority voting strategy requires them to be contained only in the majority of factors. It can handle the case of an uncorrupted element contained in estimators of all the factors except one factor that has a large amount of dense noise in the element. Therefore, the majority voting strategy is more suitable to cases with dense noise. However, with little dense noise, it will introduce some false positive elements into the estimation.

4.3 Algorithm of RMFP

In order to efficiently solve the problem in Equation (3), we propose a novel distributed robust regression algorithm, RMFP, in Algorithm 1.

In the algorithm, the active set Ψ is initialized as all the factors from 1 to M, and the initial uncorrupted set S_0 is set as the entire data samples in Line 1. Then the algorithm follows an intuitive strategy of updating $\boldsymbol{\beta}^{(m)}$ for each factor to provide a better fit for the current estimated set S in line 4, and updating the residual vector for each factor in Line 5. It then estimates the uncorrupted set $S^{(m)}$ of each factor via heuristic hard thresholding in Line 6 based on the residual vector $\boldsymbol{r}^{(m)} = \boldsymbol{y}^{(m)} - X^T \boldsymbol{\beta}^{(m)}$ in the current iteration. In Lines 7 and 8, the factor whose residual difference compared to the previous iteration is smaller than a predefined threshold will be removed from the active set Ψ. After the uncorrupted set of all the factors is computed, the estimation of all the factors will be consolidated in Line 9 with either the global consensus or majority voting strategy. The algorithm continues until the active set Ψ is empty. To improve the efficiency of the algorithm, the for loop from Lines 3 to 8 can be run in parallel for each factor.

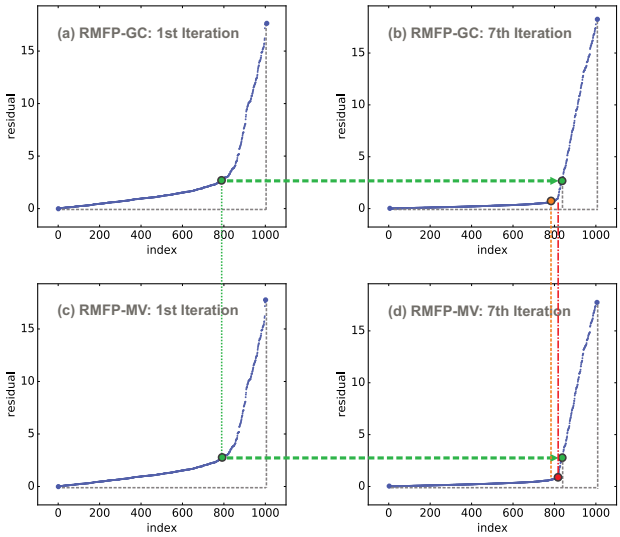


Figure 1: Residual r of one factor in ascending order for the 1st (left) and 7th (right) iterations in global consensus and majority voting strategies.

Figure 1 shows the residual of the uncorrupted set of one factor in the 1st and 7th iterations for both global consensus and majority voting strategies, respectively. It intuitively explains the convergence progress of our algorithm: The optimization steps of $\boldsymbol{\beta}^{(m)}$ based on the consolidated uncorrupted set S_t make the overall $\boldsymbol{r}_{S_t}^{(m)}$ smaller than its previous iteration; then these items in S_t have a much higher possibility to be kept in S_{t+1} than items in the set $[n] \setminus S_t$. This progress continues until the consolidated uncorrupted set is fixed. Figures 1(b) and 1(d) show that the correlated uncorrupted set by global consensus strategy has a smaller size than majority voting strategy.

5 RECOVERY ANALYSIS

In this section, the convergence analyses for the case with a global consensus strategy will be presented.

LEMMA 1. *Let τ^t be the estimated uncorrupted threshold at the t-th iteration. If $\tau_* = \gamma n$, then each factor's residual satisfies* $\|\boldsymbol{r}_{S_t}^t\|_2^2 \leq \left[1 + \frac{128(1-\gamma)}{2\gamma - 1}\right] \|\boldsymbol{r}_{S_*}^t\|_2^2$.

Lemma 1 gives an upper bound of the residual value of the estimated uncorrupted set compared to the ground truth uncorrupted set. When γ is very close to 1, $\|\boldsymbol{r}_{S_t}\|_2^2$ reaches its upper bound $\|\boldsymbol{r}_{S_*}\|_2^2$. To prove the theoretical recovery of regression coefficients, we require that the least squares function satisfies the *Subset Strong Convexity (SSC)* and *Subset Strong Smoothness (SSS)*, which are defined as follows:

DEFINITION 6. ***SSC and SSS Properties***. *The least squares function* $f(\boldsymbol{\beta}) = \|\boldsymbol{y}_S - X_S^T\boldsymbol{\beta}\|_2^2$ *satisfies the* $2\zeta_\gamma$*-Subset Strong Convexity property and the* $2\kappa_\gamma$*-Subset Strong Smoothness property if the following holds:*

$$\zeta_\gamma I \preceq \frac{1}{2}\nabla^2 f_S(\boldsymbol{\beta}) \preceq \kappa_\gamma I \quad \text{for } \forall S \in S_\gamma \tag{8}$$

Note that Equation (8) is equivalent to:

$$\zeta_\gamma \leq \min_{S\in S_\gamma} \lambda_{min}(X_S X_S^T) \leq \max_{S\in S_\gamma} \lambda_{max}(X_S X_S^T) \leq \kappa_\gamma$$

where λ_{min} and λ_{max} are denoted as the smallest and largest eigenvalues of matrix X, respectively.

THEOREM 1. *Let* $X = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n] \in \mathbb{R}^p$ *be the given data matrix and the corrupted output of each factor* $\boldsymbol{y}^{(m)} = X^T\boldsymbol{\beta}_*^{(m)} + \boldsymbol{u}^{(m)} + \boldsymbol{\varepsilon}^{(m)}$ *with* $\|\boldsymbol{u}^{(m)}\|_0 = \gamma n$*. Let* Σ_0 *be an invertible matrix such that* $\tilde{X} = \Sigma_0^{-1/2}X$*,* $f(\boldsymbol{\beta}^{(m)}) = \|\boldsymbol{y}_S^{(m)} - \tilde{X}_S\boldsymbol{\beta}^{(m)}\|_2^2$ *satisfies the SSC and SSS properties at level* α*,* γ *with* $2\zeta_{\alpha,\gamma}$ *and* $2\kappa_{\alpha,\gamma}$*. If the data satisfies* $\frac{\varphi_{\alpha,\gamma}}{\sqrt{\zeta_\alpha}} < \frac{1}{2}$*, after* $t = O\left(\log_{\frac{1}{\eta}} \frac{\|\boldsymbol{u}\|_2}{\sqrt{n}\epsilon}\right)$ *iterations, Algorithm 1 yields an* ϵ*-accurate solution* $\boldsymbol{\beta}_t^{(m)}$ *with* $\|\boldsymbol{\beta}_*^{(m)} - \boldsymbol{\beta}_t^{(m)}\|_2 \leq \epsilon + \frac{C\|\boldsymbol{\varepsilon}\|_2}{\sqrt{n}}$ *for some* $C > 0$*.*

The proof of Lemma 1 and Theorem 1 are presented in the supplementary document[1].

6 EXPERIMENT

In this section, the proposed RMFP model is evaluated on both synthetic and real-world datasets. After the experiment setup has been introduced in Section 6.1, the effectiveness of the methods is evaluated against several existing methods on both the synthetic and real-world datasets, along with an analysis of efficiency for all the comparison methods, in Section 6.2. All the experiments were conducted on a 64-bit machine with an Intel(R) core(TM) quad-core processor (i7CPU@3.6GHz) and 32.0GB memory. Details of both the source code and sample data used in the experiment can be downloaded here.[2]

6.1 Experiment Setup

6.1.1 Datasets and Labels. Our dataset is composed of synthetic and real-world data. The simulation samples were randomly generated according to the model in Equation (1) for each factor, sampling the regression coefficients $\boldsymbol{\beta}_*^{(m)} \in \mathbb{R}^p$ as a random unit norm vector. The covariance data X was independently drawn and identically distributed from $\boldsymbol{x}_i \sim \mathcal{N}(0, I_p)$, and the uncorrupted response variables were generated as $\boldsymbol{y}_*^{(m)} = \boldsymbol{X}^T\boldsymbol{\beta}_*^{(m)}$. The set of uncorrupted points S_* was selected as a uniformly random (n-τ_*)-sized subset of $[n]$, where τ_* is the size of the uncorrupted set. The corrupted response vector for each factor was generated as $\boldsymbol{y}^{(m)} = \boldsymbol{y}_*^{(m)} + \boldsymbol{u}^{(m)} + \boldsymbol{\varepsilon}^{(m)}$, where the corruption vector $\boldsymbol{u}^{(m)}$ was sampled from the uniform distribution $[-5\|\boldsymbol{y}_*^{(m)}\|_\infty, 5\|\boldsymbol{y}_*^{(m)}\|_\infty]$ and the additive dense noise was $\varepsilon_i^{(m)} \sim \mathcal{N}(0, \sigma^2)$. According to our correlated corruption assumption, the corruption vector for each factor used the unified corruption set S_*, where we have $S_* = \overline{supp(\boldsymbol{u}^{(m)})}$. Note that although each factor shares the same corruption set, the volume of corruption is factor dependent.

The real-world dataset we use is published in [37], which is built from the Sina microblog (the Chinese counterpart of Twitter). In that dataset, 1,721 volunteers, who have enough Sina microblog data, are recruited to participate in the data collection process. All of them are required to finish the Big Five questionnaire [38] online as their personality labels, and their microblogs and public personal information such as age, gender, and personal description are authorized to be obtained. We retrieve the Linguistic Inquiry and Word Count (LIWC) features from the user microblogs, which contain 63 features in total. We use the first 1,000 samples as training data and the remaining samples as testing data. Also, invalid and noisy data, such as too short answering time, is kept to evaluate our robust model.

6.1.2 Evaluation Metrics. For the synthetic data, we measured the performance of the regression coefficients recovery using the averaged L_2 error

$$e = \frac{1}{M}\sum_{m=1}^{M} \|\hat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{\beta}_*^{(m)}\|_2$$

where $\hat{\boldsymbol{\beta}}^{(m)}$ represents the recovered coefficients for each method and $\boldsymbol{\beta}_*^{(m)}$ is the true regression coefficients. To validate the performance for corrupted set discovery, precision, recall, and F1-score are measured by comparing the discovered corrupted sets with the actual ones. To compare the scalability of each method, the CPU running time for each of the competing methods was also measured.

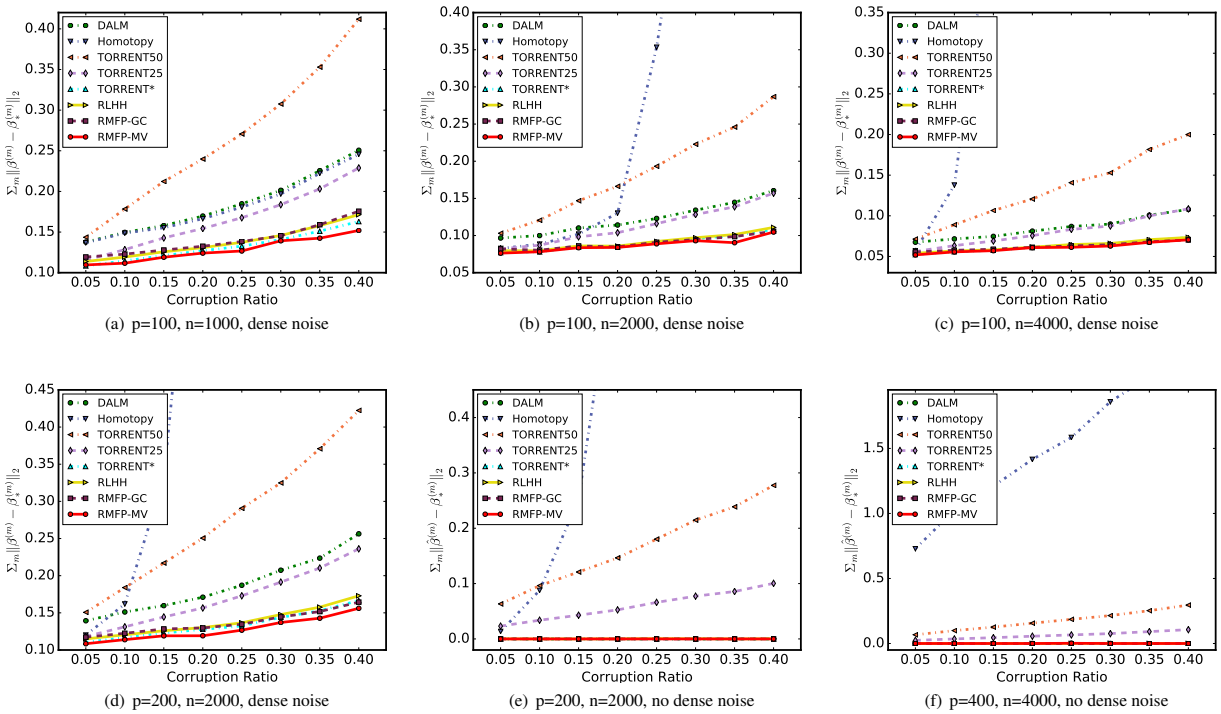
For the real-world dataset, we use the Pearson correlation coefficient (PCC) to evaluate the linear correlation between the predicted personality score $\hat{\boldsymbol{y}}^{(m)}$ and labeled personality $\boldsymbol{y}_*^{(m)}$ as follows:

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$$

where μ_X is the mean of X and σ_X is the standard deviation of X. The Pearson correlation coefficient has a value between -1 and +1, where +1 is total positive linear correlation, 0 represents no linear correlation, and -1 stands for total negative linear correlation.

6.1.3 Comparison Methods. The following methods are included in the performance comparison presented here: *Ordinary least squares* (*OLS*). The *OLS* method ignores the corruption of data and trains the model based on the whole dataset. We also compared our method to the regularized L_1 algorithm for robust regression [25] [26]. For extensive L_1 minimization solvers, [39] showed that the *Homotopy* and *DALM* solvers outperform other proposed methods both in terms of recovery properties and running time. Both of the L_1 solver methods are parameter free. A hard thresholding method, *TORRENT (abbreviated "Torr")*, developed for robust regression [29] was also compared to our method. As the method requires a parameter for the corruption ratio, which is difficult to estimate in practice, we chose three versions with different parameter settings: *TORR**, *TORR25*, and *TORR50*. *TORR** uses the true corruption ratio as its parameter, and the others apply parameters that are uniformly distributed across the range of ±25%, and ±50% off the true value, respectively. Another recently proposed heuristic hard thresholding method, *RLHH* [40], is also compared in our experiment. The method is a parameter-free approach, as it estimates the corruption set by a heuristic hard thresholding method . As all these methods are not designed for the multi-factor robust regression problem with

[1] https://goo.gl/cERzJ3
[2] https://goo.gl/JEoo5j



(a) p=100, n=1000, dense noise (b) p=100, n=2000, dense noise (c) p=100, n=4000, dense noise

(d) p=200, n=2000, dense noise (e) p=200, n=2000, no dense noise (f) p=400, n=4000, no dense noise

Figure 2: Performance on regression coefficients recovery for different corruption ratios.

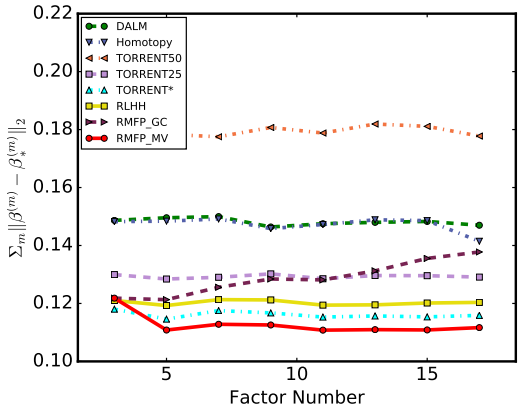


Figure 3: Regression coefficients recovery based on factor number with p=100, n=1000, and dense noise.

correlated corruption, we run them individually for each factor under the correlation corruption assumption. For our proposed methods, we use *RMFP-GC* and *RMFP-MV* to represent the *RMFP* algorithm with a global consensus or majority voting strategy, respectively. All the results will be averaged over 10 runs.

6.2 Performance

This section presents the recovery performance of the regression coefficients and the uncorrupted sets.

6.2.1 Recovery of regression coefficients. We selected seven competing methods with which to evaluate the average recovery performance of all the factors: *OLS*, *DALM*, *Homotopy*, *TORR**, *TORR25*, *TORR50*, and *RLHH*. As the recovery error for the *OLS* method is almost 10 times larger than those of the other methods, its result is not shown in Figure 2 in order to present the other results properly. Figures 2(a), 2(b), and 2(c) show the recovery performance for different data sizes when the feature number is fixed. Looking at the results, we can conclude: 1) The *RMFP-MV* method outperforms all the competing methods, including *TORR**, whose corruption ratio parameter uses the ground truth value. Also, the *RMFP-GC* method has a very competitive result compared to *TORR** and *RLHH*. 2) The results of the *TORR* methods are significantly affected by their corruption ratio parameters; *TORR50* performs almost twice as badly as *TORR** and yields worse results than one of the L_1-Solver methods, *DALM*. However, both *RMFP-GC* and *RMFP-MV* perform consistently throughout, with no impact of the parameter. 3) The L_1-Solver methods generally exhibit worse performance than the hard thresholding based algorithms. Specifically, compared to *DALM*, *Homotopy* is more sensitive to the number of corrupted instances in the data. Figure 2(d) shows its similar performance when the feature number increases. Figures 2(e) and 2(f) show that both the *RMFP-GC* and *RMFP-MV* performs equally as well as *TORR** and *RLHH* without dense noise, with both achieving an exact recovery of regression coefficients $\boldsymbol{\beta}$.

Figure 3 shows the result of regression coefficient recovery based on different factor numbers, from which we conclude: 1) The *RMFP-MV* algorithm outperforms the competing methods in all the settings of factors except when the factor is three. It is because the majority

Table 2: Precision, Recall, and F1 scores for the performance on uncorrupted set recovery.

	p=100, n=1000, o=5				p=100, n=2000, o=5			
	10%	20%	30%	40%	10%	20%	30%	40%
TORR50	0.996,0.949,0.967	0.991,0.867,0.925	0.985,0.774,0.867	0.975,0.650,0.780	0.996,0.941,0.968	0.993,0.869,0.926	0.987,0.776,0.869	0.979,0.653,0.783
TORR25	0.995,0.968,0.981	0.990,0.928,0.958	0.984,0.878,0.928	0.977,0.814,0.888	0.996,0.968,0.982	0.992,0.930,0.960	0.986,0.880,0.930	0.978,0.815,0.889
TORR*	0.994,0.994,0.994	0.987,0.987,0.987	0.980,0.980,0.980	0.971,0.971,0.971	0.995,0.995,0.995	0.989,0.989,0.989	0.982,0.982,0.982	0.972,0.972,0.972
RLHH	0.995,0.981,0.988	0.988,0.974,0.981	0.981,0.963,0.972	0.973,0.939,0.956	0.995,0.986,0.991	0.990,0.979,0.985	0.983,0.971,0.977	0.973,0.954,0.964
RMFP-GC	**1.000**,0.913,0.954	**1.000**,0.910,0.953	**1.000**,0.896,0.945	**1.000**,0.815,0.897	**1.000**,0.932,0.965	**1.000**,0.928,0.963	**1.000**,0.919,0.958	**1.000**,0.895,0.944
RMFP-MV	**1.000,1.000,1.000**	**1.000,1.000,1.000**	0.999,**1.000,0.999**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**
	p=100, n=4000, o=5				**p=200, n=2000, o=5**			
	10%	20%	30%	40%	10%	20%	30%	40%
TORR50	0.996,0.941,0.968	0.993,0.869,0.927	0.988,0.776,0.869	0.983,0.655,0.786	0.996,0.941,0.968	0.992,0.868,0.926	0.986,0.775,0.868	0.975,0.650,0.780
TORR25	0.996,0.968,0.982	0.992,0.930,0.960	0.987,0.881,0.931	0.981,0.818,0.892	0.996,0.968,0.982	0.991,0.929,0.959	0.986,0.880,0.930	0.977,0.814,0.888
TORR*	0.995,0.995,0.995	0.989,0.989,0.989	0.982,0.982,0.982	0.975,0.975,0.975	0.995,0.995,0.995	0.988,0.988,0.988	0.981,0.981,0.981	0.970,0.970,0.970
RLHH	0.995,0.987,0.991	0.990,0.984,0.987	0.983,0.974,0.978	0.976,0.966,0.971	0.995,0.984,0.990	0.989,0.978,0.983	0.983,0.964,0.973	0.973,0.944,0.958
RMFP-GC	**1.000**,0.936,0.967	**1.000**,0.935,0.966	**1.000**,0.926,0.962	**1.000**,0.921,0.959	**1.000**,0.924,0.961	**1.000**,0.922,0.959	**1.000**,0.909,0.952	**1.000**,0.876,0.933
RMFP-MV	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	0.999,**1.000,1.000**
	p=200, n=4000, o=5				**p=100, n=4000, o=5(no dense noise)**			
	10%	20%	30%	40%	10%	20%	30%	40%
TORR50	0.996,0.941,0.968	0.992,0.868,0.926	0.989,0.777,0.870	0.982,0.655,0.785	**1.000**,0.944,0.971	**1.000**,0.875,0.933	**1.000**,0.786,0.880	**1.000**,0.667,0.800
TORR25	0.996,0.968,0.982	0.991,0.929,0.959	0.988,0.882,0.932	0.981,0.817,0.892	**1.000**,0.972,0.986	**1.000**,0.938,0.968	**1.000**,0.893,0.943	**1.000**,0.833,0.909
TORR*	0.995,0.995,0.995	0.989,0.989,0.989	0.984,0.984,0.984	0.974,0.974,0.974	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**
RLHH	0.995,0.985,0.990	0.989,0.980,0.985	0.985,0.973,0.979	0.976,0.959,0.967	**1.000**,0.988,0.994	**1.000**,0.988,0.994	**1.000**,0.986,0.993	**1.000**,0.987,0.993
RMFP-GC	**1.000**,0.931,0.964	**1.000**,0.924,0.960	**1.000**,0.922,0.960	**1.000**,0.914,0.955	**1.000**,0.921,0.959	**1.000**,0.935,0.966	**1.000**,0.903,0.949	**1.000**,0.910,0.953
RMFP-MV	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**
	p=200, n=4000, o=3				**p=200, n=4000, o=9**			
	10%	20%	30%	40%	10%	20%	30%	40%
TORR50	0.996,0.941,0.968	0.993,0.869,0.927	0.988,0.776,0.869	0.982,0.655,0.786	0.996,0.941,0.968	0.993,0.869,0.926	0.988,0.777,0.870	0.982,0.654,0.785
TORR25	0.996,0.968,0.982	0.992,0.930,0.960	0.987,0.881,0.931	0.981,0.817,0.892	0.996,0.968,0.982	0.992,0.930,0.960	0.987,0.881,0.931	0.981,0.817,0.892
TORR*	0.995,0.995,0.995	0.990,0.990,0.990	0.982,0.982,0.982	0.974,0.974,0.974	0.995,0.995,0.995	0.989,0.989,0.989	0.982,0.982,0.982	0.974,0.974,0.974
RLHH	0.995,0.986,0.991	0.991,0.981,0.986	0.983,0.972,0.978	0.976,0.957,0.966	0.995,0.986,0.991	0.990,0.981,0.985	0.984,0.971,0.977	0.976,0.959,0.967
RMFP-GC	**1.000**,0.959,0.979	**1.000**,0.959,0.979	**1.000**,0.949,0.974	**1.000**,0.938,0.968	**1.000**,0.882,0.937	**1.000**,0.880,0.936	**1.000**,0.855,0.922	**1.000**,0.842,0.914
RMFP-MV	**1.000,0.999,0.999**	0.999,**1.000,0.999**	0.998,**0.999,0.998**	0.997,**0.998,0.998**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**	**1.000,1.000,1.000**

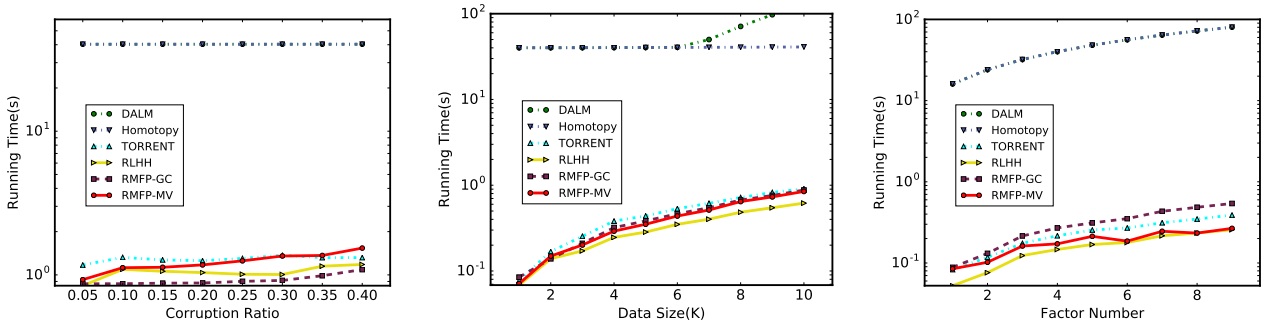
number of three factors is so small that it is easier to include the corrupted samples and exclude uncorrupted ones than a larger factor number. 2) The recovery error of the *RMFP-GC* algorithm increases when the number of factors rises. It is because the correlation corruption property used in *RMFP-GC* restricts the uncorrupted samples are accepted by all the factors, which leads to more uncorrupted samples being excluded from the estimation. 3) Similar to the result in Figure 2, hard thresholding based algorithms generally outperform L_1-Solver methods in different settings of factor numbers. However, the *TORR50* algorithm performs worse than other methods because a badly estimated corruption ratio is used as its parameter.

6.2.2 Recovery of Uncorrupted Sets. As most competing methods do not explicitly estimate uncorrupted sets, we compared our proposed method with the *RLHH* and *TORR* algorithms using a number of different parameter settings ranging from the true corrupted ratio up to a deviation of 50%. Based on the results in different settings of features, samples, and factors (denoted by *p*,*n*, and *o* respectively), shown in Table 2, we found: 1) The *RMFP-MV* method outperforms all the competing methods with very high precision, recall, and F1 scores. It is important to note that *RMFP-MV* even performs better than *TORR** with the true corruption ratio parameter, which cannot be estimated exactly in practice. This indicates that our method can significantly improve the result of uncorrupted set estimation in the correlated corruption property. 2) Because of the global consensus design, the *RMFP-GC* method achieves the highest precision in retrieving the uncorrupted samples. However, since the correlated corruption property is too strictly followed, the method also excludes some uncorrupted samples from the estimation set, which leads to a lower value of recall compared to *RMFP-MV*. 3) The results of the *TORR* methods are highly dependent on the corruption ratio parameter: The results for a true corruption estimation error are much better than those for 25% and 50% errors. However, the *RMFP* algorithm is a parameter-free method that is capable of consistently obtaining a good result. 4) When increasing the feature number and corruption ratio, we obtain a similar performance. Specifically, *RMFP-MV* can obtain all the scores closed to 1.00 in different data settings, while the performance of other methods degrades when the corruption ratio increases. 5) In different settings of factor numbers, we conclude that *RMFP-MV* performs worse when the factor number decreases. It is mainly because less correlated corruption information can be used in smaller factor numbers. However, the recall of *RMFP-GC* increases in three factors since fewer factors are used to constrain the uncorrupted estimation as compared to five factors. 6) In a no-dense-noise setting, the *RMFP-MV* method performs an optimal recovery result, while *TORR** exactly recovers the result because it is using the true corruption ratio.

6.2.3 Result of Personality Prediction. To evaluate the performance of personality prediction, we compared our proposed methods of *RMFP-GC* and *RMFP-MV* to competing methods, including

Table 3: Pearson Correlation of Personality Prediction

	Agr.	*Con.*	*Ext.*	*Ope.*	*Neu.*	**Avg**
OLS	0.2461/35.46%	0.2437/33.07%	**0.1921/24.95%**	0.2733/37.03%	0.0910/11.88%	0.2092/28.48%
TORR	0.2075/29.90%	0.2157/29.27%	0.1766/22.94%	0.2405/32.59%	0.0971/12.68%	0.1875/25.47%
RLHH	0.2111/30.42%	0.2332/31.64%	0.1867/24.25%	0.2739/37.11%	0.1064/13.89%	0.2023/27.46%
RMFP-GC	0.2146/30.92%	0.2296/31.15%	0.1887/24.51%	0.2552/34.58%	**0.1098/14.33%**	0.1996/27.10%
RMFP-MV	**0.2472/35.62%**	**0.2442/33.13%**	0.1919/24.92%	**0.2743/37.17%**	0.0967/12.62%	**0.2109/28.69%**



(a) p=400, n=4000, no dense noise (b) p=100, cr=0.1, dense noise (c) p=100, cr=0.1, dense noise

Figure 4: Running time for different corruption ratios and data sizes

OLS, *RLHH*, and *TORR* algorithm. For the setting of the *TORR* algorithm, a fixed 10% corruption ratio is used as its parameter, while the majority number in the *RMFP-MV* algorithm is set as 4 out of 5. The *Pearson correlation coefficient* is used as a metric to evaluate the correlation between the estimated personality scores and labeled scores. The result is shown in Table 3, where the columns represent the scores of the five personality factors: *agreeableness*(*Agr.*), *conscientiousness*(*Con.*), *extraversion*(*Ext.*), *openness*(*Ope.*), *neuroticism*(*Neu.*), and their average scores. As one person can get varied personality scores in different personality questionnaires, we use the correlation score from a test-retest reliability experiment [41] as the optimal value in this domain. The value is estimated by correlating scores obtained in the first test with scores from a second test approximately six week later. The optimal correlation value for each factor in the Chinese language [42] is listed as follows: *Agr.* (0.694), *Con.* (0.737), *Ext.* (0.770), *Ope.* (0.738) and *Neu.* (0.766). In Table 3, we show the percentage of the evaluated method compared to the optimal value in the second column of each factor.

From the result in Table 3, we conclude: 1) *RMFP-MV* outperforms all the competing methods in an average of the Pearson correlation. However, the result of *RMFP-GC* is 5.6% worse than *RMFP-MV* since it uses a more strict assumption of correlated corruption. 2) The *RLHH* method, which considers the robustness for each factor independently, only competes with the *OLS* method in two factors: *Ope.* and *Neu.* The facts show that the correlated corruption property applied in *RMFP-MV* can improve overall performance for multiple factors. 3) The *TORR* algorithm using a fixed corruption ratio performs worse than other methods because some corrupted samples can be included if the estimated corruption is less than the actual corruption. 4) Compared to other factors, the performance of *Neu.* is 55.8% worse on average for all the methods, which means the *Neu.* factor cannot be properly handled by evaluated methods in the collected data. Also, we found that average correlation can be improved by 12.56% if the *Neu.* factor is removed.

6.2.4 Efficiency. To evaluate the efficiency of our proposed method, we compared the performance of all the competing methods for three different data settings: different corruption ratios, data sizes, and factor numbers. In general, as Figure 4 shows, we conclude: 1) The *RMFP* method has a very competitive performance even though it performs the additional consolidation step in either global consensus or majority voting in each optimization iteration. The efficiency difference compared to *RLHH* and *TORR* is trivial, which indicates that the consolidation step in *RMFP* always performs efficiently in different data settings. 2) The running time for *RMFP* methods increases slowly as either the data size or factor number increases, just as in the *TORR* and *RLHH* methods. When the corruption ratio increases, the running time of *RMFP* increases within 10%, which means that data corruption has little impact on the efficiency of our method. 3) The hard thresholding based methods significantly outperform the L_1-Solver based methods.

7 CONCLUSION

In this paper, a novel robust model is proposed to handle multi-factor personality prediction in the presence of correlated corruption. To achieve this, we designed a heuristic hard thresholding method to estimate the corruption set along with global consensus or majority voting strategies, that is alternately updated with the optimized regression coefficients. We demonstrate that our algorithm can handle a general multi-factor robust regression problem in the property of correlated corruption with a strong recovery guarantee on regression coefficients in a geometric convergence rate. Extensive experiments on both synthetic data and real-world data demonstrated that the proposed algorithm outperforms other comparable methods in both effectiveness and efficiency.

REFERENCES

[1] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transaction on Affective Computing*, 5(3):273–291, 2014.

[2] Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, 1992.

[3] Kelly Moore and James C. McElroy. The influence of personality on facebook usage, wall postings, and regret. *Computers in Human Behavior*, 28(1):267 – 274, 2012.

[4] George Domino and Marla L Domino. *Psychological testing: An introduction*. Cambridge University Press, 2006.

[5] Deniz S Ones and Chockalingam Viswesvaran. The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human performance*, 11(2-3):245–269, 1998.

[6] Richard N. Landers and John W. Lounsbury. An investigation of big five and narrow personality traits in relation to internet usage. *Comput. Hum. Behav.*, 22(2):283–293, March 2006.

[7] Adam W Meade and S Bartholomew Craig. Identifying careless responses in survey data. *Psychological methods*, 17(3):437, 2012.

[8] Matthias Ziegler, Carolyn MacCann, and Richard Roberts. *New perspectives on faking in personality assessment*. Oxford University Press, 2011.

[9] Emily S Orr, Mia Sisic, Craig Ross, Mary G Simmering, Jaime M Arseneault, and R Robert Orr. The influence of shyness on the use of facebook in an undergraduate sample. *CyberPsychology & Behavior*, 12(3):337–340, 2009.

[10] Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. Inferring the socioeconomic status of social media users based on behaviour and language. In *European Conference on Information Retrieval*, pages 689–695. Springer, 2016.

[11] Xiaowei Zhang, Li Cheng, and Tingshao Zhu. Robust multivariate regression with grossly corrupted observations and its application to personality prediction. In *Proceedings of The 7th Asian Conference on Machine Learning*, pages 112–126, 2015.

[12] Teresa Correa, Amber Willard Hinsley, and Homero Gil De Zuniga. Who interacts on the web?: The intersection of usersâĂŹ personality and social media use. *Computers in Human Behavior*, 26(2):247–253, 2010.

[13] Gregory J Meyer, Stephen E Finn, Lorraine D Eyde, Gary G Kay, Kevin L Moreland, Robert R Dies, Elena J Eisman, Tom W Kubiszyn, and Geoffrey M Reed. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128, 2001.

[14] Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL '15, pages 1754–1764, 2015.

[15] Glen Coppersmith Mark Dredze Craig Harman. Quantifying mental health signals in twitter. *ACL 2014*, 51, 2014.

[16] Daniel Preotiuc-Pietro, Maarten Sap, H Andrew Schwartz, and LH Ungar. Mental illness detection at the world well-being project for the clpsych 2015 shared task. *NAACL HLT 2015*, page 40, 2015.

[17] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. *Icwsm*, 11(1):281–288, 2011.

[18] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[19] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[20] Fabio Celli, Elia Bruni, and Bruno Lepri. Automatic personality and interaction style recognition from facebook profile pictures. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1101–1104. ACM, 2014.

[21] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle H Ungar. Analyzing personality through social media profile picture choice. In *ICWSM*, pages 211–220, 2016.

[22] A. Smolic and J. R. Ohm. Robust global motion estimation using a simplified m-estimator approach. In *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, volume 1, pages 868–871 vol.1, 2000.

[23] Peter J. Huber and Elvezio M. Ronchetti. *The Basic Types of Estimates*, pages 45–70. John Wiley & Sons, Inc., 2009.

[24] Yoonsuh Jung, Seung Pil Lee, and Jianhua Hu. Robust regression for highly corrupted response by shifting outliers. *Statistical Modelling*, 16(1):1–23, 2016.

[25] John Wright and Yi Ma. Dense error correction via l1-minimization. *IEEE Trans. Inf. Theor.*, 56(7):3540–3560, July 2010.

[26] Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via l1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.

[27] Yiyuan She and Art B. Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

[28] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In Sanjoy Dasgupta and David Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 774–782. JMLR Workshop and Conference Proceedings, May 2013.

[29] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[30] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1503–1512, New York, NY, USA, 2015. ACM.

[31] Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pages 733–742, Arlington, Virginia, United States, 2010. AUAI Press.

[32] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.

[33] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 109–117, New York, NY, USA, 2004. ACM.

[34] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[35] Seyoung Kim and Eric P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In Johannes FÃijrnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 543–550. Omnipress, 2010.

[36] Matthias Ziegler and Markus Buehner. Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, 69(4):548–565, 2009.

[37] Xiaoqian Liu and Tingshao Zhu. Deep learning for constructing microblog behavior representation to identify social media userâĂŹs personality. *PeerJ Computer Science*, 2:e81, September 2016.

[38] Gian Vittorio Caprara, Claudio Barbaranelli, Laura Borgogni, and Marco Perugini. The âĂIJbig five questionnaireâĂİ: A new questionnaire to assess the five factor model. *Personality and Individual Differences*, 15(3):281 – 288, 1993.

[39] Allen Yang, Arvind Ganesh, Shankar Sastry, and Yi Ma. Fast l1-minimization algorithms and an application in robust face recognition: A review. Technical Report UCB/EECS-2010-13, EECS Department, University of California, Berkeley, Feb 2010.

[40] Zhang Xuchao, Zhao Liang, Boedihardjo Arnold P., and Lu Chang-Tien. Robust regression via heuristic hard thresholding. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, IJCAI'17. AAAI Press, 2017.

[41] Samuel D Gosling, Peter J Rentfrow, and William B Swann. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

[42] Richard Carciofo, Jiaoyan Yang, Nan Song, Feng Du, and Kan Zhang. Psychometric evaluation of chinese-language 44-item and 10-item big five personality inventories, including correlations with chronotype, mindfulness and mind wandering. *PLOS ONE*, 11(2):1–26, 02 2016.