A Preliminary Study on Cleaning up Erroneous Data and Filling in Missing Values in A Medical Record

Er-Wei Bai, 
Department of Electrical and Computer Engineering, University of Iowa, Iowa City, Iowa 52242, er-wei-bai@uiowa.edu

Hans Johnson, Weiyu Xu, and Mathews Jacob

Abstract

This study investigates various possible approaches in dealing with missing or erroneous data sets. Cleaning up erroneous values and filling in missing values in a meaningful way play a critical role for the consequent data analysis and decision making. What makes a meaningful way depends on available prior information. In this preliminary study, a number of possible approaches are proposed and tested on an actual Huntington disease data set, depending on the assumptions.

Keywords

missing data; erroneous data; Gaussian process model; low rank matrices

1. INTRODUCTION

For a variety of reasons, a medical data set often has missing and erroneous values. Data missing can be due to a record or human error. Erroneous data can be from measurement noise, device malfunction or operator failure. Even in a normal situation, different machine readings or different interpretations by clinical staffs can cause some deviations in a medical record. For instance, MRI results from different clinics have various bias due to different settings, and also individual clinical staff may interpret results differently. Some may interpret conservatively and others are not.

Those missing and erroneous values could have a big impact for the consequent modeling and analysis. Simply discarding an entire or a big part of record because the values of one variable is missing, useful even critical information can be lost. For erroneous values, even the first question of how to detect if a value is erroneous is non-trivial. Unfortunately, in most data modeling and analysis, one starts with an assumption that the data is accurate. Clearly without cleaning up the data or filling in missing values in a meaningful way, the consequently analysis and conclusions become questionable. In this paper, we study several types of missing and erroneous data and propose ways to deal with the problems. The results are tested on a real Huntington disease data set.

2. PROBLEM STATEMENT

Let the true (but unknown) data matrix be given by
This representation has a number of applications. For instance the columns of the matrix could indicate features or attributes and the rows represent subjects or diseases or categories, e.g., \(q_{ij}\) is the jth feature of the ith subject for a particular disease. It is often the case that some features are correlated linearly or nonlinearly and the matrix \(Q\) is a low or approximately low (in terms of singular values) rank matrix.

In practice, the true \(Q\) is unavailable. Assume that the available data matrix is given by

\[
P = \begin{pmatrix}
p_{11} & \cdots & p_{1n} \\
\vdots & \ddots & \vdots \\
p_{m1} & \cdots & p_{mn}
\end{pmatrix} = (p_{ij}).
\]

Let \(M_k, k = 1, 2\) be two subsets of

\[
\{(i, j), i = 1, \ldots, m, j = 1, \ldots, n\}
\]

We consider two cases.

I. **Missing Data:**
   a. \(P = Q\). However, for \((i, j) \in M_1\), the values \(q_{ij} = p_{ij}\) are missing, where \(M_1\) is assumed to be known. In other words, \(Q\) is available except for some missing elements \(q_{ij}, (i, j) \in M_1\). The question is how to recover these missing values.
   b. \(P = Q\). However, for \((i, j) \in M_1\), the values \(q_{ij} = p_{ij}\) are missing. Further \(Q\) is assumed to be a low rank matrix, \(r = \text{rank } Q < \min(n, m)\).

II. **Missing data in the presence of outliers:**
   \(P = Q + O\), where the value \(q_{ij}\) is missing for \((i, j) \in M_1\) and \(o_{ij}\)'s are outliers that \(\neq 0\) if and only if \((i, j) \in M_2\). \(M_1\) is assumed to be known but not \(M_2\) which is assumed to be sparse in the sense that the number of non-zero elements of \(M_2\) is much much smaller than the total number of elements \(n \times m\) of \(M_2\). \(Q\) could be a low rank matrix.

3. **POSSIBLE SOLUTIONS**

The goal is to find the true but unknown \(q_{ij}\)'s from \(p_{ij}\)'s in some optimal sense under the assumptions.
3.1 Case I

This is basically a nonlinear non-parametric prediction/estimation problem. There exists a number of ways that can be modified for our application.

Without loss of generality, let \( y_{i0} = (q_{i1}, \ldots, q_{il}) \) be a missing element vector in the \( ith \) row of \( Q \) and \( Y_{i} = (q_{i(l+1)}, q_{i(l+2)}, \ldots, q_{i(n-l)}) \) be an available data vector. Then, the question is what is the best estimate of \( y_{i0} \) given \( Y_{i} \) as well as \( y_{j0}, Y_{j}, j \neq i \). We study this problem in a statistical setting. Suppose the distributions of \( q_{ij}, j = 1, 2, \ldots, n \)’s are not independent and further the conditional density of \( y_{i0} \)

\[ f(y_{i0}|Y_{i}) \]

conditioned on \( Y_{i} \) is available. Then, if the minimum variance estimate is of concern, the best estimate of \( \hat{y}_{i0} \) of \( y_{i0} \) is the conditional mean. With this idea, we consider a Gaussian process model (Rasmussen et al (2006)). Suppose \( y_{i0} \) and \( Y_{i} \) are jointly Gaussian with mean

\[ \begin{bmatrix} E_{y_{i0}} \\ E_{Y_{i}} \end{bmatrix} \]

and a covariance matrix

\[ \Sigma = \begin{bmatrix} A & B \\ B' & C \end{bmatrix} \]

It follows that \( y_{i0} \) and \( Y_{i} \) follow a joint Gaussian

\[ \begin{bmatrix} y_{i0} \\ Y_{i} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} E_{y_{i0}} \\ E_{Y_{i}} \end{bmatrix}, \begin{bmatrix} A & B \\ B' & C \end{bmatrix} \right) \]

The conditional density of the missing \( y_{i0} \) conditioned on the available data vector \( Y_{i} \) is also Gaussian

\[ y_{i0} \sim \mathcal{N}(E_{y_{i0}} + BC^{-1}(Y_{i} - E_{Y_{i}}), A - BC^{-1}B') \]

The minimum variance estimate \( \hat{y}_{i0} \) of \( y_{i0} \) is the conditional mean given by

\[ \hat{y}_{i0} = E_{y_{i0}} + BC^{-1}(Y_{i} - E_{Y_{i}}). \quad (3.1) \]

Now further assume that rows of \( Q \) follow a joint Gaussian distribution. Then, the values of the matrices \( EY_{i}, A, B \) and \( C \) can be calculated from rows of \( Q \) where no data is missing so called the training data in the literature.
3.3 Case II

In this case, $Q$ is assumed to be a low rank matrix and $q_{ij}$'s, $(i, j) \in M_1$ are missing. Consider the following optimization problem

$$\min \| \hat{Q} \|_* \quad \text{subject to } \hat{q}_{ij} = q_{ij}, \ (i, j) \not\in M_1$$

(3.2)

where $\| \cdot \|_*$ stands for the nuclear norm of a matrix (Candes et al. (2012)). Let $r = \text{rank } Q$. $M$ be the number of entries of $Q$ where the value is not missing and $w = \max(n, m)$. As derived in (Candes et al. (2012)) (Theorem 6.1 in Appendix), under some technical conditions, the solution $\hat{Q}$ of the minimization (3.2) recovers every missing (and non-missing) element of $Q$ exactly with a high probability (Candes et al. (2012)) provided $M > cw^{1.2} r \log w$. In short, as long as $Q$ is low rank and there are enough non-missing entries in terms of the rank and dimension of $Q$, the missing values can be accurately recovered.

3.3 Case II

Here $P = Q + O$. Suppose $O$ is sparse in the sense that the number of non-zero elements is much much smaller than the number of zero elements of $O$ (Wright et al. (2009)). Let

$$\hat{Q} = (\hat{q}_{ij}) \text{ and } \hat{O} = (\hat{o}_{ij})$$

Consider

$$\min \| \hat{O} \|_1 + \lambda \| \hat{Q} \|_* \quad \text{s.t. } \hat{Q} + \hat{O} = P$$

(3.3)

where $\| \cdot \|_1$ is the $l_1$ matrix norm. From (Wright et al. (2009)) (Theorem 6.2 in Appendix), $Q$ and $O$ can be exactly recovered if $O$ is sparse. Interested readers may find details from (Wright et al. (2009)).

4. TEST EXAMPLE

The test data was a real Huntington disease data from (Paulsen et al. (2014)). The size of the data matrix $Q \in \mathbb{R}^{61 \times 8}$ is 61 by 8. Eight columns of the data matrix $Q$ contains information of Neurological Exam results (Column 1), Cognitive Tests 1–6 (Columns 2–7) and Motor Evaluations (Column 8) respectively. Each row of $Q$ represents a subject. The value of $Q$ is shown in Figure 1.

To form $P = Q + O$, 10 outliers were added independently at random entries of the data matrix $O$. Each added outlier was a Gaussian variable of zero mean and variance $50^2$. The exact indices and values of $O$ were
The corrupted data matrix \( P \) is shown Figure 2. To recover \( Q \) from \( P \), we applied the proposed method (3.3) in Case II. For \( \lambda = 1 \), the recovered \( \hat{Q} \) is shown in Figure 3. To show how close between \( Q \) and \( \hat{Q} \) in the presence of significant outliers, we calculate the relative errors between the true but unknown \( Q \) and the recovered \( \hat{Q} \) element by element, \( \frac{|q_{ij} - \hat{q}_{ij}|}{q_{ij}} \), \( i = 1, \ldots, 61 \), \( j = 2, \ldots, 7 \) for the columns 2–7. In Figure 4, each line represents \( \frac{|q_{ij} - \hat{q}_{ij}|}{q_{ij}} \), \( i = 1, \ldots, 61 \) for some \( j \in [2, 3, 4, 5, 6, 7] \). Note the first and the last column contains zero and relative error does not make sense. The majority error is within 10% percent.

Next, we assumed that the values of \( \hat{q}_{ij} \), \( i = 41, \ldots, 61 \), \( j = 1, 2, 3 \) of \( \hat{Q} \) were missing as shown in Figure 5. In others words, the neurological exam results and cognitive test results 1–2 of the last 21 subjects were missing. The question was could we recover or estimate these missing values from the remaining cognitive test results 3–6 and motor evaluation results of the last 21 subjects as well as the complete results of the first 40 subjects? To this end, the Gaussian process model (3.1) was applied. The first 40 subjects data was used as a training data to calculate \( EY = EY, Ey_{0} = Ey_{0} \) and \( A, B, C \). Then from the available data \( Y_{i} = (q_{i4}, q_{i5}, q_{i6}, q_{i7}, q_{i8})' \), \( i = 41, \ldots, 61 \)

the missing values \( y_{i0} = (q_{i1}, q_{i2}, q_{i3})' \), \( i = 41, \ldots, 61 \)

were calculated by (3.1). The original values and recovered values are in Figure 6, where Gaussian process model of (3.1) was used. The true but unknown values are the last three columns and the estimated values are the first three columns. To make comparisons, the low rank matrix completion method of (3.2) was used and the result is shown in Figure 7. Again, the true but unknown values are the last three columns and the estimated values are the first three columns. The low rank matrix completion method (3.2) seems to work better in this case.

5. CONCLUDING REMARKS

In this preliminary study, we investigate an important but often neglected issue, i.e., how to deal with a data set with missing and erroneous values. A number of possible methods are proposed depending on assumptions. From the preliminary study, the proposed methods seem to work with the given Huntington disease data. The reason behind is that the data matrix \( Q \) is indeed low or approximately low rank as illustrated in Figure 8 where the singular values \( \sigma_{1}^{2} \geq \ldots \geq \sigma_{8}^{2} \) of \( Q \)
are shown. Clearly, there is one and only one dominant singular value. Therefore, Q is approximately low rank. We believe that cleaning up the data prior to analysis is certainly a direction that deserves more attentions.

References

1. Candes EJ, Recht B. Exact matrix completion via convex optimization. Comm of ACM. 2012; 55:111–119.
2. Rasmusse CE, Williams C. Gaussian Process for Machine Learning. The MIT Press; Cambridge, MA: 2006.
3. Paulsen J, Long J, et al. Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of predict-HD study. Frontiers in aging neuroscience. 2014; 6:1–11. Article 78. [PubMed: 24478697]
4. Wright J, Peng Y, Ma Y, Ganesh A, Rao S. Ro- bust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. In: Bengio Y, Schumans D, Lafferty J, Williams C, Culotta A, editors in Advances in Neural Information Processing Systems. Vol. 22. 2009.

APPENDIX

The following two results are from (Candes et al (2012)) and (Wright et al (2009)) respectively.

Theorem 6.1

Let $Q \in \mathbb{R}^{n \times m}$ be of rank $r$ sampled from orthogonal model and let $l = \max(n, m)$. Suppose we observe $w$ entries of $Q$ with locations sampled uniformly at random. Then there exist two constants $c_1, c_2 > 0$ such that if

$$w \geq c_1 l^{15/4} r \log l$$

the optimization of (3.2) is unique and equal to $Q$ with probability at least $1 - c_2 l^{-3}$.

Write the singular value decomposition of $Q \in \mathbb{R}^{n \times m}$,

$$Q = U \sum \sigma_i u_i v_i^* = \sum_{i=1}^r \sigma_i u_i v_i^*.$$

Then, the incoherence condition with parameter $\mu$ states that

$$\max_i \|u_i^* e_i\|^2 \leq \frac{\mu r}{n} \max_i \|v_i^* e_i\|^2 \leq \frac{\mu r}{m} \|Uv_i^*\|_\infty \leq \sqrt{\frac{\mu r}{nm}}$$

where $e_i$ is a unite vector with $ith$ entry being unity.
Fig. 1.
The true but unknown Huntington data set Q. Eight columns of the data matrix Q contains information of Neurological Exam results (Column 1), Cognitive Tests 1–6 (Columns 2–7) and Motor Evaluations (Column 8) respectively. Each row of Q represents a subject.

Theorem 6.2. Suppose the support set of $O$ is uniformly distributed among all sets of cardinality $w$. Then, there exist constants $c_1, c_2, c_3 > 0$ such that with probability at least $1 - c_1 n^{-10}$, the optimization (3.3) recovers $Q$ with $\lambda = 1/\sqrt{n}$, provided that $r \leq c_2 m \mu^{-1} (\log n)^{-2}$ and $w \leq c_3 nm$. 

---

Proc IFAC World Congress. Author manuscript; available in PMC 2018 August 27.
Fig. 2.
The corrupted data matrix $P = Q + O$. 

---

*Proc IFAC World Congress. Author manuscript; available in PMC 2018 August 27.*
Fig 3.
The recovered Q.
Fig. 4.
Relative estimation errors $|q_{ij}^* - q_{ij}|/q_{ij}$, $i = 1, \ldots, 61$, columns $j = 2, \ldots, 7$. 
The values of the first three columns of the last 21 subjects are missing.
Figure 6.
Missing and recovered values by Gaussian process models. The last three columns are true but missing and the first three columns are recovered values.
Fig. 7.
Missing and recovered values by low rank matrix approach. The last three columns are true but missing and the first three columns are recovered values.
Fig 8.
Singular values of the data matrix $Q$. 

Proc IFAC World Congress. Author manuscript; available in PMC 2018 August 27.