MISSING DATA RECOVERY FOR FARM FINANCIAL DATA: AN APPLICATION OF THE MATRIX COMPLETION METHOD

by

YU WANG

(Under the Direction of Jeffrey H. Dorfman)

ABSTRACT

The Agricultural Resource Management Survey (ARMS) dataset is source of information on production and financial practice of farm businesses and the farm households in U. S. However, one variable missing from ARMS data is commodity-specific returns. While revenue is recorded on a commodity-specific basis, input usage data are generally not, meaning that commodity-specific returns can only be computed for single-commodity producing farms. Possessing commodity-specific return distributions for a sample of U.S. farms would be very useful for agricultural policy analysis. In this dissertation, I utilize a matrix completion approach to recover the missing commodity-specific net return values in ARMS dataset and estimate the suitable fitted distributions for those net returns for six major commodities. Overall, the matrix completion approach is efficient at recovering these “missing” values and allows policy makers access to highly useful information.

INDEX WORDS: Matrix Completion, Missing Data Recovery, Farm Financial Data

MISSING DATA RECOVERY FOR FARM FINANCIAL DATA: AN APPLICATION OF THE MATRIX COMPLETION METHOD

by

YU WANG

B.A., NORTHWEST A & F UNIVERSITY, CHINA, 2012

M.S., NORTHWEST A & F UNIVERSITY, CHINA, 2014

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2018

MISSING DATA RECOVERY FOR FARM FINANCIAL DATA: AN APPLICATION OF THE MATRIX COMPLETION METHOD

by

Yu Wang

Major Professor:	Jeffrey H. Dorfman
Committee:	Levi A. Russell
	Chen Zhen

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2018

DEDICATION

I dedicate this dissertation to my wife, Linlin Guo, for her great love, considerate support and encourage during all these years we have been through together; and to my parents, Wenqian Wang and Shulian Wang, who always encourage me to overcome the difficulties in life and give me endless love since my childhood.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Jeffrey H. Dorfman. Since I was admitted by the department, he gave me endless support, valuable guidance, considerate care and warm patience every day. You not only offer me wise advice on my research, but also guide me to be a thoughtful, rigorous and smart person like you. I have tremendous respect and appreciation to you. It is so lucky and fortunate for me to be your student during the time in University of Georgia. You are the best advisor for a Ph. D who can meet. I will remember the days spending with you forever.

I would also like to thank my committee members, Dr. Levi A Russell and Dr. Chen Zhen for their great support and value advices. I have been benefitted a lot by their encouragement, humorous words and inspiration to pursue a Ph. D degree in United States. They demonstrate how can become an experienced and dedicated expert in Agricultural Economics.

During the time I stay in University of Georgia, a lot of friends offer their warm and kind help to me. I would like to thank all of my friends, especially to Xiaoxiao Sun, Zhaochong Liu, Yibo Dang, Xiaohan Mei and Shiyu Ye. I will always remember the enjoyable days spending with you. Thank you so much for encouraging and supporting me. I will memorize the year accompanying with all of my good friends forever.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

American agriculture plays an important role in global agricultural markets and government agricultural policy has a profound influence on U.S. agricultural markets. When crafting agricultural policy, how the policies impact the level and distribution of farm income is a critical consideration. Policymakers have access to plenty of data on farm income from a whole-farm perspective, but data on commodity-specific earnings is much scarcer, which can cause difficulty in formulating agricultural policies when features that are more commodity-specific are considered.

An essential information source to make policy in American agriculture sector is the Agricultural Resource Management Survey (ARMS) database. The ARMS database contains information on income, cost, productivity, demographic information, resource allocation and other financial conditions of American farmers and farm businesses. The ARMS database can allow us to forecast farm financial information such as net farm incomes and study farm production practices. However, what is missing from ARMS is information on commodity-specific returns. It would be very useful for agricultural policy analysis if such information could be precisely estimated from the ARMS data.

The revenue questions in ARMS are commodity-specific questions, but the input usage data are generally not. This means commodity-specific returns can only be computed for single-commodity producing farms. ARMS provides no commodity-specific net returns in general. The

research goal of this dissertation is to estimate returns on each commodity produced by the farmers based on ARMS data. To accomplish this goal, the net returns of each commodity are regarded as missing values.

Thus, the central research question is: is it possible for us to make precise estimates of the "missing" values of commodity-specific returns? This question belongs to a famous class of questions called the Netflix Prize Problem, which is nearly the same as ARMS missing-data dilemma. In Netflix's recommendation system, each user submits his/her ratings on a subset of Netflix films and tv shows that they have watched and chosen to rate. Netflix uses these partial ratings to predict likely ratings for each user to thousands of other viewing options. The limited range of ratings means that many missing values occur in Netflix's ratings data matrix. It is desirable for Netflix to complete the data matrix so that Netflix can recommend movies to any users who are willing to order films. Netflix's prize offer to researchers who could help solve this missing values problem led to the general research problem of matrix completion being referred to as the Netflix Problem.

This dissertation will show that the missing commodity-specific net returns for farmers in the ARMS dataset can be recovered using matrix completion methods and that the resulting distribution of returns on commodities is accurate enough to use in the policy making process. This dissertation will proceed to do this by first reviewing some literature, then presenting the methodology, and finally by demonstrating the method on data with known returns using data from Texas A&M's Agricultural and Food Policy Center and then on ARMS data.

CHAPTER 2

LITERATRUE REVIEW

Mathematically, we can regard the ARMS data problem as following: There is a matrix $\boldsymbol{W}$ with n_1 rows and n_2 columns. The indices of rows represent farmers and the indices of columns represent different commodities. We only observe *w* entries in this matrix, where *m* is much smaller than $n_1 \times n_2$, the total number of entries. We wish to estimate all the missing values in $\boldsymbol{W}$ based on those *m* entries. There are several ways to impute the missing values in matrix $\boldsymbol{W}$. Economists have been most likely to use Bayesian methods to impute missing values, so we will explore that literature first, then the more mathematically focused matrix completion algorithms.

There are three categories of missing data in a Bayesian framework: Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR). MCAR data means that the missing value is not related to any observed value or other missing values. MAR data indicates that the missing value only depends on some observed values. NMAR data stands for missing values related to some unclear functions generated by observed data or missing data.

People always want to find some parameters, α or β, to make inference on missing data with the help of these parameters. α or β can denote the mean or variance of the data. α or β can also indicate the coefficient in the regression. Different choices of parameters depend on the different model specification and different kinds of missing data. To impute the missing values, analysts first divide the whole dataset M into two groups: observed data (M_{obs}) and missing data (M_{miss}).

Analysts assume that the distribution of missing data follows the posterior distribution which is proportional to the prior distribution of the observed data. We denote the set of parameters as δ. By integrating out the parameters based on the observed data, people can get the probability distribution of the parameter as:

$$p(\delta|M_{obs}) = \int p\,(\delta|M)p(M_{miss}|M_{obs})dM_{miss} \quad (2.1)$$

Equation (2.1) holds when the data is MAR. If we define the missing indices matrix as N, then $N_{ij} = 1$ when M_{ij} is in M_{miss} and $N_{ij} = 0$ when M_{ij} is in M_{obs}. We assume the distribution of N is $p(N|M, \theta)$ with a given value of M and unknown parameter θ. Combing the assumed statistical model and incomplete dataset together, we can derive the joint likelihood of δ and θ as (Rubin, 1976):

$$L(\delta, \theta|M_{obs}, N) \propto \int p(M_{miss}, M_{obs}|\delta)\, p(N|M_{miss}, M_{obs}, \theta)dM_{miss} \quad (2.2)$$

Based on the likelihood function (2.2), we can derive the joint distribution δ and θ as (Mitra, 2008):

$$p(\delta, \theta|M_{obs}, N) \propto p(\delta, \theta) \int p(M_{miss}, M_{obs}|\delta)\, p(N|M_{miss}, M_{obs}, \theta)dM_{miss} \quad (2.3)$$

where $p(\delta, \theta)$ is the prior distribution and $p(\delta, \theta) = p(\delta)p(\,\theta)$. Applying the joint distribution of parameters to impute the missing values is defined as Bayesian Proper Imputation (Rubin, 1987). Since the joint likelihood and joint distribution are difficult to compute, analysts use the Expectation-Maximization (EM) algorithm to find the maximum expected value of the log-likelihood derived based on the whole dataset. This EM algorithm at T step is defined as (Little, 2011):

$$L(\delta|M_{obs}, \delta^t) \propto \int \log p(M_{miss}, M_{obs}|\delta)\, p(M_{miss}|M_{obs}, \delta = \delta^t)dM_{miss} \quad (2.4)$$

Then at T+1 step, the new values of missing data and parameters can be drawn as:

$$M_{miss}^{t+1} \sim p(M_{miss}|M_{obs}, \delta^t) \quad (2.5)$$

$$\delta^{t+1} \sim p(\delta | M_{obs}, M_{miss}^{t+1}) \quad (2.6)$$

In conclusion, the basic strategy under Bayesian Proper Imputation is to draw the missing values and parameters and replace them in the next step until the log-likelihood function achieves the maximum value. It is quite similar to the Gibbs sampler process.

Another method to impute missing values under Bayesian framework is Multiple Imputation. Analysts create q datasets denoted as $D^t = (M_{obs}, M_{miss}^t)$ for $t = 1, 2, \ldots, q$, where M_{miss}^t are drawn from posterior distribution $p(M_{miss} | M_{obs}, \delta^t)$ at the t step (Zhou and Reiter, 2010). Then define the estimated distribution of δ as

$$\epsilon = \lim_{q \to \infty} \frac{1}{q} \sum_{t=1}^{q} \int p(\delta | M_{miss}^t, M_{obs}) d\delta \quad (2.7)$$

The total amount of simulated δ values from $p(\delta | D^t)$ for each D^t is T. The corresponding δ value at each time is denoted as $\widehat{p(\delta^t)}$. Then the analysts mix and sort all the values $\widehat{p(\delta^t)}$. The $(\epsilon q T)$th element among all the sorted values is the estimate of δ.

Under a Bayesian framework, the missing data and the parameters are all random. We only know there should exist prior and posterior distributions for these unknown data and parameters. However, it is impossible to figure out the specific model which defines the missing data generation mechanism. Another difficulty is that there is no information on the covariates of the distribution for the parameters. In brief, we have no information about the statistical properties of the missing data. We need to define the model and prior distribution by ourselves. However, such a process requires skills to translate subjective beliefs into a mathematical formulation. Recovering the missing values accurately will depend highly on selecting the right model and prior distribution. Additionally, the posterior distributions are heavily influenced by the priors. Since different people have different choices about models and prior distributions, we can't ensure that the recovery effect is good when the model and prior distribution change. Another

disadvantage of Bayesian approach to imputing missing values is the computational cost. As we explained in the previous section, we need to draw imputed missing values and estimated parameters from the posterior distribution a lot of times. The ARMS data matrix is quite large, the computational cost is very intractable if we use a Bayesian approach to impute the missing values in every column. In addition, there are three classifications of missing data when applying the Bayesian method to impute missing values. We don't know which kind of missing data is the best fit in ARMS dataset. It is nearly impossible for us to choose right model and prior distribution to impute these missing data.

The second method is hot deck imputation. The basic idea of hot deck imputation is to replace the missing values by the observed values. The missing values are denoted as "non-respondent" or "recipient". The observed values are denoted as "respondent" or "donor". There are two approaches to impute missing values when applying hot deck imputation (Andridge, R. R., and Little, R. J., 2010). The first approach is "random hot imputation". It indicates that missing values are imputed by choosing values from the set of donors randomly. The second version is "deterministic hot imputation". The chosen values from donor set based on certain standard. The key tasks in hot deck imputation are to create donor set and match the donors to the recipients.

A simple way to adjust the donor set to imputed cell is based on subjective knowledge. We can justify whether the property of the donor cells is also held by the missing values. If we are sure that the property of missing value is the same as that of donor set, we can choose from the donor set to replace missing value. A more general way to match the donors to recipients is using some statistical measurement. If $T = (T_{u1}, T_{u2}, \ldots T_{uv})$ denotes v observed values on subject u and j denotes the subject of missing values, the closeness $d(u, j)$ between the observed values and missing values can be defined as (Andridge, R. R., and Little, R. J., 2010):

$$d(u,j) = \max\left|T_{uv} - T_{jv}\right| \tag{2.8}$$

Also, we can use the difference between the predicted value of observed and missing data:

$$d(u,j) = \max\left|\hat{Y}(T_u) - \hat{Y}(T_j)\right| \tag{2.9}$$

where $\hat{Y}(T_u) = \hat{\beta}T_u$ and $\hat{\beta}$ is coefficient estimation derived by only regressing on the observed data. After we get the imputed values from the donor set, we should consider assigning different weights to the imputed values from the different subjects. Rao and Shao (1992) points out that we can select sample weights based on the sample size of each subject. In addition, we can assign equal weights to different subjects if there are a lot of equal predicted values for these subjects.

We should notice that hot deck imputation does not rely on model fitting for the variable to be imputed and is potentially less sensitive to model misspecification compared to Bayesian approach. However, there is no explicit statement on the choice of statistical measurement when matching the donors to the recipients. Through hot deck imputation, the imputed values are very similar to the observed values. In some extreme cases, all the imputed values in subject *j* are equal to the values in subject *u* one by one. This is because the imputed values are all chosen from the observed data. Another drawback of hot deck imputation is that we can't make useful statistical inferences as we can with either the Bayesian approach and a matrix completion approach. For example, we have no idea about the degrees of freedom or the p-values in the specified model. Sometimes, hot deck imputation can be regarded as a naïve method to replace the missing values by the observed data.

Matrix completion is the third potential way to recover missing values in the ARMS database. For a matrix completion approach, we can use very minimal storage space. The computational cost of each-iteration is very low as the matrix is recovered very quickly. On the theoretical side,

a sequence of iterates converges in matrix completion. Thus, we don't need to know the specific prior distribution and model specification. The matrix completion approach is amenable to large-scale problems by recovering the low-rank matrix only based on a small amount of observed entries. To apply a matrix completion approach, we construct a matrix $\boldsymbol{W}$ which consists of the desired variables related to financial information of the farmers. By minimizing the nuclear norm of $\boldsymbol{W}$, we can recover the original ARMS dataset exactly. More details about applying matrix completion approach are illustrated in the methodology part.

A variety of methods for recovering a low-rank matrix with many missing values based on a small number of observations are known collectively as matrix completion approaches. Such approaches most famously have been applied to solve the Netflix Prize Problem and other similar problems (Candès and Tao, 2009). Candès and Recht (2009) prove that we can recover an incomplete square matrix with a high probability if the number of sampled elements, m, satisfies the condition that $m \geq Cn^{1.2}r\log n$, where *C* is a positive numerical constant, *n* denotes the number of rows and r is the rank of this matrix. Candès and Tao (2009) propose that any unknown matrix can be recovered exactly by nuclear norm minimization when an assumption on the singular vectors of this matrix and a condition on the order of information theoretic limit hold. Keshavan et al. (2009) derive an algorithm to compute the singular value decomposition of $n\alpha \times n$ matrix $\boldsymbol{M}$ based on the observed element set $|E| = O(r\ n)$. The stopping criteria is defined by the root mean square error $RMSE \leq C(\alpha)(\frac{nr}{|E|})^{\frac{1}{2}}$. However, the optimal RMSE would decay with $|E|/nr$. There is still room for improvement for more general models of exact matrix completion.

There is also a further consideration about completing an unknown matrix with noise only based on a few sampled entries. Candès and Plan (2010) point out that nuclear norm

minimization can help recover a low-rank matrix based on $nrlog^2n$ noisy elements sampled with error, where n is the number of rows in matrix, r is the rank of the matrix and number of error is proportional to the total number of elements. Keshavan and Montanari (2010) propose a cost function minimized by the OPTSPACE algorithm when the noise level is very large and difficult to capture by the signal model. Klopp (2014) defines two nuclear-norm penalized estimators to optimize the convex problem under high-dimensional scaling without any information on variance or mean of the noise if the noise distributed following a general sampling distribution.

As matrix completion has attracted a lot of attention, researchers have developed several approaches with quite different algorithms for solving equivalent problems. Candès and Tao (2009) derive a minimum number of sampled elements to recover a matrix containing a large proportion of missing values based on convex optimization programming. They propose a new singular value decomposition algorithm to recover the low-rank square matrix with rank r. Recht (2011) points out that the sampling entries are always assumed to be under a Bernoulli model in previous work. He proposes a new incoherence condition on the sampled entries of a low-rank matrix from the uniform model and then recovers this matrix exactly under singular value decomposition. Nathan Srebro et al. (2005) investigate such a matrix completion problem by applying a Maximum-Margin Matrix Factorization (MMMF) method. By minimizing trace norm of a binary target matrix using linear programming and factorizing the target matrix into two semi-definite matrices, we can get an optimal solution matrix, the specific recovered entries and predictions for the new observations. Cai et al. (2008) develop a Singular Value Thresholding (SVT) algorithm by adding a Frobenius-Norm term to the optimization function. At each step, the algorithm sets up a theoretical thresholding on the singular values of a transition matrix and

produce a sequence of solution matrices and transition matrices. The optimal solution matrix is found when the Frobenius norm of the solution matrix is maximized under the thresholding. Ryan Kennedy (2013) proposes a gradient descent method and compute the missing values in one column by an updating process with other columns fixed first. He also compares the effects of different methods to recover a low-rank matrix when missing data proportions and singular values change. Hastie et al. (2014) combine the MMMF and SVD approaches together to derive an algorithm for large and sparse matrix factorization and completion, which they called the **softImpute-ALS** algorithm. By constraining the rank of the solution matrix during the MMMF process, Hastie et al. derive a solution matrix recovering the low-rank target matrix exactly within a high dimensional space.

With different efficient algorithms to recover large and sparse matrices, the matrix completion approach can now be applied to different areas. Cai et al. (2015) first propose a new matrix completion algorithm to recover missing values under the framework of structured matrix completion. This approach can recover a low rank matrix when there is only one missing block in the matrix. The authors apply this method to genomic data to investigate missing miRNA values. Candès et al. (2015) apply a matrix completion approach to recover diffracted images. By solving a convex objective function, they can recover signal data with noise-free measurement under one or two dimensions. Wang et al. (2011) construct a nonparametric version of denoising swissroll data with a matrix completion framework. Building on Manifold Blurring Mean Shift (MBMS), they apply a matrix completion algorithm to 100-Dimension data and Mocap data to check the efficiency of this algorithm. Athey et al. (2017) apply the matrix completion approach to a panel dataset to measure a causal effect by minimizing the nuclear norm of the difference between an objective matrix and the approximation matrix. Combining

their estimator with two different patterns of missing data, they can extend their approach into a fixed effects model by adding time series structure among the missing data.

CHAPTER 3

MTHODOLOGY

Candès and Tao (2009) propose that it is very possible for us to get precise numeric value of missing elements in a data matrix $\boldsymbol{W} \in \mathbb{R}^{n_1 \times n_2}$ and, thus, to gain an overall picture of whole data matrix. We follow the procedures and work from Candès and Tao (2009), Candès and Recht (2008), Cai et al. (2008) and Hastie et al. (2015) to present how their theoretical and methodoligical innovations can be applied to solve our dilemma. We only observe *w* entries from matrix $\boldsymbol{W}$, where w is a small proportion of the total elements of the matrix $\boldsymbol{W}$. The rank of matrix $\boldsymbol{W}$ is *k* with dimension $n_1 \times n_2$, where $k \ll \min(n_1, n_2)$. The low value of *k* indicates that the matrix $\boldsymbol{W}$ is a low rank matrix. In our problem, the matrix $\boldsymbol{W}$ is the ARMS dataset. n_1 is the number of rows (farmers) and n_2 is the number of columns (variables) in the dataset. We define $n \equiv \max(n_1, n_2)$. The observed statistics of farmers in ARMS dataset, w_{ij}, can be expressed by the location indicator (i, j) $\in \Omega$, where Ω is the complete location set corresponding to matrix $\boldsymbol{W}$. There are $[n_1] \times [n_2]$ total number of entries in this location set. Based on Candès and Tao's (2009) work, the prerequisite condition to prove why this theory is valid in our application is that the data matrix $\boldsymbol{W}$ has approximately low rank. This means that we only observe a little fraction of statistics for all of farmers. Based on the common knowledge of ARMS data, this prerequisite condition holds in our case. However, we should make some basic concepts clear and set up necessary assumptions if we want to recover the ARMS dataset $\boldsymbol{W}$ as accurately as possible.

An important concept associated with recovering the matrix $\boldsymbol{W}$ is the singular value decomposition (SVD). The basic idea of a singular value decomposition is to factor $\boldsymbol{W}$ into the

product of three matrices, $\boldsymbol{W} = UDV^T$ for any $\boldsymbol{W}$, where U and V are orthonormal matrices and D is diagonal matrix with positive real entries. We are interested in whether we can get all the information from target $\boldsymbol{W}$ or not when degrees of freedom of this matrix is $(n_1 n_2 - k)k$. The property of descriptive statistics in U and V is that they are highly related to the degrees of freedom of $\boldsymbol{W}$.

Suppose the available elements in matrix $\boldsymbol{W}$ can be denoted as w_{ij}, where (i, j) ∈ Ω. If the two location indicators are unique in Ω, the element observed or recovered are unique because the ARMS data matrix, $\boldsymbol{W}$, is a two-dimensional data matrix. In location set Ω, we always assume that every indicator can be distributed uniformly and chosen randomly to denote an observation. There should exist an orthogonal projection matrix of Ω, P_Ω, which is projected onto all matrices in Ω orthogonally. Suppose the real number set $\boldsymbol{K}$ is $\mathrm{P}_\Omega(\boldsymbol{Q})$, where $\boldsymbol{K}_{ij} = \begin{cases} \boldsymbol{Q}_{ij}, & \text{if (i, j)} \in \Omega \\ 0, & otherwise \end{cases}$. If $\boldsymbol{Q}_{ij} = \boldsymbol{W}_{ij}$, it indicates that $\boldsymbol{W}_{ij}$ is observed in the location set. As all the information in these two matrices are the same, $\boldsymbol{Q}$ is a perfect substitute for $\boldsymbol{W}$. Considering the relationship between one matrix and its projection, we know that $\mathrm{P}_\Omega(\boldsymbol{W}) = \mathrm{P}_\Omega(\boldsymbol{Q})$. It is also reasonable for us to make inferences about $\boldsymbol{W}$ from the projection matrix $\mathrm{P}_\Omega(\boldsymbol{W})$ and $\mathrm{P}_\Omega(\boldsymbol{Q})$ because they share same orthogonal space. Now, we know basic idea in order to recover the ARMS data matrix, $\boldsymbol{W}$, when it is a low-rank target matrix: derive its corresponding recovered matrix under with the help of its solution matrix $\boldsymbol{Q}$ and its projection matrix.

Candès and Recht (2008) point out that if one can find a simplest-form matrix fitting the observed entries constrained by the rank, such a matrix is the most efficient candidate to recover target matrix $\boldsymbol{W}$. Thus, one should look for a fitting matrix containing the smallest possible total number of elements. Furthermore, if all the columns or rows in this fitting matrix are independent from each other, this matrix must hold the most succinct formulation according with

the observed elements. We know that the number of pivot elements indicates the number of independent rows or columns in the matrix. Mathematically, the rank of a matrix is equal to number of pivot elements. Such an approach can be transformed into the following optimization problem if we want to look for such a matrix in the simplest formulation (Candès and Tao, 2009):

$$\text{minimize: rank } (\mathbf{Q}) \tag{3.1}$$

$$\text{subject to } \boldsymbol{Q}_{ij} = \boldsymbol{W}_{ij}, (\text{i}, \text{j}) \in \Omega$$

Under the previous analysis, the upper limit for the dimension of the matrix **Q** is *n*, which is equal to the dimension of $\boldsymbol{W}$. It is inefficient to find the minimum rank of **Q** because we should show proof work about dependent or independent relationship between every column and row in **Q**. In linear algebra, a nuclear norm can be thought of as the relaxation of numbers of non-zero eigenvalues, which is equal to the rank of a matrix. The nuclear norm of a matrix is denoted as: $\|\boldsymbol{Q}\|_* = \sum_i \sigma_i(\boldsymbol{Q})$. Based on the work of Candès and Tao (2009), equation (1) can be expressed as:

$$\text{minimize: } \|\boldsymbol{Q}\|_* \tag{3.2}$$

$$\text{subject to } \mathrm{P}_\Omega(\boldsymbol{Q}) = \mathrm{P}_\Omega(\boldsymbol{W})$$

A potential impediment is that we don't how many elements are zeros in the matrix. For instance, if there exist a matrix $\boldsymbol{M}$ which is defined as $\begin{bmatrix} 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$, it is impossible to recover it. The only recovered values we get from $\boldsymbol{M}$ are repeated zeros. Following Candès and Recht's (2008) structure, we present a lower bound for the minimum number of observed elements available in the ARMS data matrix $\boldsymbol{W}$ in order to be able to recover missing elements. The two authors propose the minimal sampling requirement theorem on such lower bound. If the

observed statistics w_{ij} are chosen randomly and uniformly based on the location indicators i and j, there should exist a numerical constant C satisfying (Candès and Tao, 2009)

$$w \geq Cn^{5/4}k\log(n) \tag{3.3}$$

Given the condition in (3), there will be a unique decision matrix $\boldsymbol{Q}$. The probability for recovering $\boldsymbol{W}$ by $\boldsymbol{Q}$ is at least $1 - cn^{-3}$ for some constant c. Considering that C can be any numerical constant, equation (3) is intuitive to show that we can always recover an unknown matrix with a lot of missing values as long as the number of observed entries in target matrix is not too small. This is a very important idea for recovering a data matrix like the ARMS dataset, which is a quite large data matrix with high percentage of missing values.

To illustrate our problem in detail, we describe Candès and Recht's (2008) work for proving minimal sampling in following section.

3.1 Minimal Sampling Requirement

Before developing some results on minimal sampling requirements, a few more details are useful about the Singular Value Decomposition (SVD) method to factorize any matrix, including the target matrix $\boldsymbol{W}$. These results hold for any matrix. The decomposition of $\boldsymbol{W}$ expressed by singular vectors can be defined as: $\boldsymbol{W} = UDV^T = \sum_{1\leq j\leq k} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^*$, where U and V are orthogonal matrices, D is a diagonal matrix with positive real entries, σ_j is the singular value (the root of the eigenvalue of $\boldsymbol{W}^*\boldsymbol{W}$) and $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are the basis element spanned in U and V. Orthogonal matrices mean that $U'U = V'V = I$. If subspace U belongs to $\mathbb{R}^n$ with dimension r, P_U would be the orthogonal projection onto U. Under the SVD framework for matrix completion, the accuracy in recovering a target matrix $\boldsymbol{W}$ depends on the singular vectors and basis element chosen in subspace U and V. We need to avoid a large proportion of basis element being in the null space corresponding to U and V.

Candès and Recht (2008) define the coherence of U as:

$$\mu(U) \equiv \frac{n}{k} max \|\mathrm{P}_U \boldsymbol{e}_i\|^2 \text{ for } 1 \leq i \leq n. \tag{3.4}$$

$\mu(U)$ will achieve maximum as n/k and minimum as 1. Low coherence indicates that few standard basis elements are in the null space of any subset of U or V. We also know that standard basis is uncorrelated to the singular vector when coherence is low. If the standard basis elements or the singular vectors can spread widely except they are on the null space, they can contain all information from the projection matrices. This is a very important property if we want to recover matrix $\boldsymbol{W}$ much more precisely. To prohibit the wide spread of singular vectors and basis elements which we want to avoid, Candès and Recht (2008) propose two important assumptions for minimal sampling requirements to be met.

Assumption $\boldsymbol{A_0}$: For specific positive μ_0, coherence should follow the inequality

$\max(\mu(U), \mu(V)) \leq \mu_0$

Assumption $\boldsymbol{A_1}$: For some positive μ_1, every element in matrix $\sum_{1 \leq j \leq k} \boldsymbol{u}_j \boldsymbol{v}_j^*$ with dimension $n_1 \times n_2$ is constrained by a maximum value, which is $|\mu_1 \sqrt{k/(n_1 n_2)}|$.

Under Assumptions $\boldsymbol{A_0}$ and $\boldsymbol{A_1}$, we show how the minimal sampling theorem applies in our case, following the same pattern described by Candès and Recht's (2008).

Recall that ARMS dataset $\boldsymbol{W}$ is $n_1 \times n_2$ matrix with rank k and $n = \max(n_1, n_2)$. Based on the location indicators from their original orthogonal space, we observe some elements of $\boldsymbol{W}$ chosen uniformly and randomly. There should exist some constants C and c satisfying that if

$$w \geq C max(\mu_1^2, \mu_0^{1/2} \mu_1, \mu_0 n^{\frac{1}{4}}) nk\beta \log(n) \tag{3.5}$$

for some $\beta \geq 2$, then we can get unique recovered matrix for $\boldsymbol{W}$ with probability at least $1 - cn^{-3}$. When the rank is very small, we can get recover $\boldsymbol{W}$ with high probability provided that

$$w \geq C\mu_0 n^{6/5} k\beta \log(n) \tag{3.6}$$

Combining equation (3.5) and (3.6), we get the rough idea about minimal sampling theorem.

With this background, we will show the minimal sampling theorem is valid step by step. With the help of assumption $\boldsymbol{A_0}$ and $\boldsymbol{A_1}$, we relax the requirement for the number of minimal sampled elements from the target $\boldsymbol{W}$. Equation (1) can be rewritten as (Candès and Recht, 2008):

$$\text{minimize: } ||\boldsymbol{Q}||_* \tag{3.7}$$

$$\text{subject to } R(\boldsymbol{Q}) = t$$

R is a sampling operator which selects observed elements in $\boldsymbol{Q}$ randomly and combines the chosen elements into linear relationship within real number set t. The dimension of t is less than $n_1 \times n_2$. Since we already set up $\boldsymbol{Q}_{ij} = \boldsymbol{W}_{ij}$ as a one-to-one injective equation in (3.1), the constraint in equation (3.7) can be regarded as building up a linear relationship between elements in the objective matrix and the real number set. When matrix $\boldsymbol{Q}$ contains many non-zero singular values σ_k, we know:

$$||\boldsymbol{Q}||_* \leq rank(\boldsymbol{Q})||\boldsymbol{Q}\,||_{\boldsymbol{F}}^{\boldsymbol{2}} \tag{3.8}$$

by the Cauchy-Schwarz inequality.

The initial research goal is to minimize $rank(\boldsymbol{Q})$, and it can be shown that the Frobenius norm and nuclear norm of $\boldsymbol{Q}$ are minimized when we get the smallest rank of $\boldsymbol{Q}$ based on equation (3.7) and (3.8). Following work of Candès and Recht (2008), we can get the transition matrix to $\boldsymbol{W}$, $\boldsymbol{K}$, with the following properties:

$$P_T(\boldsymbol{K}) = \textstyle\sum_{1\leq j\leq k} \boldsymbol{u}_j \boldsymbol{v}_j^* \;\&\; \|P_{T^\perp}(\boldsymbol{K})\| < 1 \tag{3.9}$$

where P_T is a linear sampling operator mapping from matrix to another matrix. Define the notation: $\boldsymbol{\Lambda} = P_T - p^{-1}P_T P_\Omega P_T$. We adopt Candès and Recht's (2008) approach: for any matrix $\mathbf{Q}$ in set $\boldsymbol{T}$, $(P_T P_\Omega P_T)^{-1}(\mathbf{Q})$ can be approximated by a Taylor series expansion:

$$(P_T P_\Omega P_T)^{-1}(\mathbf{Q}) = p^{-1}(\mathbf{Q} + \boldsymbol{\Lambda}(\mathbf{Q}) + \boldsymbol{\Lambda^2}(\mathbf{Q}) + \cdots) \tag{3.10}$$

Equation (3.10) will always hold regardless of the number of dimensions contained by the target matrix. If we substitute $\boldsymbol{K}$ into equation (10), we can show that $\boldsymbol{K} = P_\Omega P_T (P_T P_\Omega P_T)^{-1} (\sum_{1 \le j \le k} \boldsymbol{u}_j \boldsymbol{v}_j^*)$, $P_{T^\perp}(\boldsymbol{K})$ can be expressed as:

$$P_{T^\perp}(\boldsymbol{K}) = p^{-1}(P_{T^\perp} P_\Omega P_T)\left(\sum_{1 \le j \le k} \boldsymbol{u}_j \boldsymbol{v}_j^* + \boldsymbol{\Lambda}\left(\sum_{1 \le j \le k} \boldsymbol{u}_j \boldsymbol{v}_j^*\right) + \boldsymbol{\Lambda^2}\left(\sum_{1 \le j \le k} \boldsymbol{u}_j \boldsymbol{v}_j^*\right) + \cdots\right) \quad (3.11)$$

By taking the limit of equation (3.10) and (3.11), we can bound the norms of the candidate matrix to recover the ARMS data matrix $\boldsymbol{W}$ in location set and the orthogonal space location set. This means that we have a finite range in which to search for a candidate matrix. Since we already have location indicators of observed entries in candidate matrix, we can derive the property of these observed entries further. With the help of the following lemmas proposed by Candès and Recht (2008), we can show the minimal sampling requirement is valid in our matrix completion framework.

Lemma 1. For $\beta \ge 2$ and $\lambda \ge 1$, there should exist a constant C_0 if $\log(n)\lambda\tau_1^2 nk\beta \le w$, we can conclude that

$$C_0 \lambda^{\frac{1}{2}} \ge p^{-1} || (P_{T^\perp} P_\Omega P_T) \sum_{1 \le j \le k} \boldsymbol{u}_j \boldsymbol{v}_j^* || \quad (3.12)$$

with probability at least $1 - n^{-\beta}$.

Lemma 2. For $\beta \ge 2$ and $\lambda \ge 1$, there should exist a constant C_1 and c_1 if $\log(n)\lambda\tau_1 nk\beta max(\tau_1, \sqrt{\tau_0}) \le w$, we can conclude that

$$C_1 \lambda^{-1} \ge p^{-1} || (P_{T^\perp} P_\Omega P_T) \boldsymbol{\Lambda}\left(\sum_{1 \le j \le k} \boldsymbol{u}_j \boldsymbol{v}_j^*\right) || \quad (3.13)$$

with probability at least $1 - c_1 n^{-\beta}$.

Lemma 3. For $\beta \ge 2$ and $\lambda \ge 1$, there should exist a constant C_2 and c_2 if $\log(n)\lambda\tau_0^{4/3} nk^{4/3}\beta n \le w$, we can conclude that

$$C_2\lambda^{-3/2} \geq p^{-1}||(P_{T^\perp}P_\Omega P_T)\boldsymbol{\Lambda^2}\left(\sum_{1\leq j\leq k}\boldsymbol{u}_j\boldsymbol{v}_j^*\right)|| \quad (3.14)$$

with probability at least $1 - c_2 n^{-\beta}$

Lemma 4. For $\beta \geq 2$ and $\lambda \geq 1$, there should exist a constant C_3 and c_3 if $\log(n)\lambda\tau_0^2 nk^2\beta n \leq w$, we can conclude that

$$C_2\lambda^{-1/2} \geq p^{-1}||(P_{T^\perp}P_\Omega P_T)\boldsymbol{\Lambda^3}\left(\sum_{1\leq j\leq k}\boldsymbol{u}_j\boldsymbol{v}_j^*\right)|| \quad (3.15)$$

with probability at least $1 - c_3 n^{-\beta}$.

Lemma 5. If we define the location set Ω and n= max (n_1, n_2), the coherence should obey the inequality, max $(\mu(U), \mu(V)) \leq \tau_0$. There should exist constant C_R, where $C_R\sqrt{\frac{\tau_0 nk\beta\log(n)}{w}} < 1$, for all $\beta \geq 1$, we can conclude that

$$C_R\sqrt{\frac{\tau_0 nk\beta\log(n)}{w}} \geq p^{-1}||(P_{T^\perp}P_\Omega P_T - pP_T|| \quad (3.16)$$

with the probability at least $1 - n^{-\beta}$.

Lemma 6. Under Lemma 5, there should exist a constant C_{K1},

log(n)$k\beta n\tau_0(2C_R)^2 \leq w$, we can conclude that

$$C_{K1}\left(\frac{\log(n)k\beta n\tau_0}{w}\right)^{k1/2}\left(\frac{n^2 k}{w}\right)^{1/2} \geq p^{-1}||(P_{T^\perp}P_\Omega P_T)\sum_{k\geq k1}\boldsymbol{\Lambda^k}\left(\boldsymbol{u}_j\boldsymbol{v}_j^*\right)|| \quad (3.17)$$

with probability at least $1 - n^{-\beta}$.

With all these lemmas, if $\tau_0 k \leq n^{3/4}$ and $(\tau_0 k)^{3/4} \leq \tau_0 n^{5/4} k$, the target matrix $\boldsymbol{W}$ can be perfectly recovered with the probability at least $1 - n^{-\beta}$ if the minimal sampling obeys the condition

$$w \geq Cmax({\tau_1}^2, \sqrt{\tau_0}\tau_1, \tau_0 n^{1/4})\log(n)nk\beta. \quad (3.18)$$

Candès and Recht (2008) point out that when the rank of matrix $\boldsymbol{W}$ is very small, the minimal sampling obeys that

$$w \geq Cmax(\tau_0{}^2, \tau_0 n^{1/5})\log(n)nk\beta \quad (3.19)$$

with probability $1 - cn^{-\beta}$. All the proofs for these lemmas is done by Candès and Recht, and are included in the appendix.

After we show that the existence of the minimum number set of observed entries from target matrix $\boldsymbol{W}$ is the sufficient condition to recover matrix $\boldsymbol{W}$, the next step is to prove that the solution $\boldsymbol{Q}$ is unique. After relaxing the constraint, we can use the Singular Value Decomposition (SVD) to express $\boldsymbol{Q}$. Under equation (3.9), we know the solution matrix $\boldsymbol{Q}$ is an unique solution to equation (3.2). To prove the uniqueness of $\boldsymbol{Q}$**,** our strategy is to make use of properties held by $\boldsymbol{K}$. $\boldsymbol{K}$is vanishing in the complementary set of location set of $\boldsymbol{Q}$. Two important properties held by $\boldsymbol{K}$ are: duality and injectivity. Duality refers to the involution theorem in math. If we know the dual of B is A, then the dual of A is B. When the involution set contains different fixed points, the dual of A is A itself. In our case, matrix $\boldsymbol{K}$ is vanishing outside the location set so that all the points in $\boldsymbol{K}$ would decay to a set of fixed points. We can conclude that if we can find a matrix consisting of fixed points, this matrix is dual to itself. Injectivity is a one-to-one definition. Injectivity can map the element in the domain of one matrix to one distinct element in the codomain of this matrix. Based on the duality, we can find a matrix dual to itself. Under injectivity, we can find accurate one-to-one relationship between two matrices. Combine the two properties together, we can find the dual matrix is the same matrix as itself, which indicates that such a matrix is unique.

Now, we can show how to find such dual matrix step by step under the work of Candès and Recht (2008) and Candès and Tao (2009). Candès and Recht (2008) rewrite the problem as:

$$\text{Minimize: } ||\boldsymbol{Q}||_F \quad (3.20)$$

$$\text{subject to } (P_T P_\Omega)(\boldsymbol{Q}) = \sum_{j=1}^{k} u_j v_j^*$$

Matrix $\boldsymbol{Q}$ will achieve the minimum Frobenius norm if it can be expressed as a constraint in equation (3.20). Candès and Recht (2008) apply the Pythagorian formula and derive that

$$\|\boldsymbol{Q}\|_F^2 = \|P_T(\boldsymbol{Q})\|_F^2 + \|P_{T^\perp}(\boldsymbol{Q})\|_F^2 = \left\|\sum_{j=1}^{k} u_j v_j^*\right\|_F^2 + \|P_{T^\perp}(\boldsymbol{Q})\|_F^2$$

$$= k + \|P_{T^\perp}(\boldsymbol{Q})\|_F^2 \tag{3.21}$$

The rank k is fixed, so $\|\boldsymbol{Q}\|_F^2$ is minimized when $\|P_{T^\perp}(\boldsymbol{Q})\|_F$ is minimized under the condition that $P_T(\boldsymbol{Q}) = \sum_{j=1}^{k} u_j v_j^*$. Compared to the Frobenius norm, the nuclear norm is more amenable to calculate. The condition in equation (21), $P_T(\boldsymbol{Q}) = \sum_{j=1}^{k} u_j v_j^*$, is very close to the form of the nuclear norm. Thus, we can set up constraints on nuclear norms of $\|P_{T^\perp}(\boldsymbol{Q})\|$ and $\|P_T(\boldsymbol{Q})\|$ to look for minimization values instead of using the Frobenius norm. We should always keep in mind that our strategy is look for an optimal matrix $\boldsymbol{Q}$ firstly, not making imputations on the real data matrix, the ARMS dataset, directly.

Following Candès and Recht's (2008) work, we present two operators:

1. Operator is $R_{\Omega T}(\boldsymbol{W}) = P_\Omega P_T(\boldsymbol{W})$

2. $\boldsymbol{K}$ is equal to $R_{\Omega T}(R_{\Omega T}^* R_{\Omega T})^{-1}(\sum_{j=1}^{k} u_j v_j^*)$

These two operators show the relationship between a unique minimizer to the ARMS dataset $\boldsymbol{W}$ and $\boldsymbol{W}$ itself. To make operators reliable, $R_{\Omega T}^* R_{\Omega T}$ should be equal to $P_T P_\Omega P_T$. $R_{\Omega T}^* R_{\Omega T} = P_T P_\Omega P_T$ shows that injective operator R maps from T to $\mathbb{R}^{n_1 \times n_2}$. There are two theorems to prove injectivity and the uniqueness of matrix $\boldsymbol{Q}$ by exploring the properties of R (Candès and Recht, 2008, proof is in appendix).

Theorem 1. If matrix $\boldsymbol{Q}$ is well defined, R is an operator getting sample elements uniformly and $n = \max(n_1, n_2)$, there should exist a numerical constant C satisfying the condition

$$p^{-1}\|P_T P_\Omega P_T - pP_T\| \leq C\sqrt{\frac{\mu n k(\log n\beta)}{w}} \tag{3.22}$$

with some probability if $C\sqrt{\frac{\mu nk(\log n\beta)}{w}} \leq 1$.

During the proof of theorem 1, the constraint on the spectral norm of operator $P_\Omega P_T$ is shown. In linear algebra, the dual norm of the nuclear norm is its corresponding spectral norm since norms often come in dual pairs if we attach a constraint to them. Theorem 1 here implies that the operator $R_{\Omega T}$ can lead to the dual of matrix $\boldsymbol{Q}$ when it is applied. Theorem 1 is also called the Rudelson selection estimate (Candès and Tao, 2009). No matter what the numerical value of μ is, we can always derive w satisfying the constraint in Theorem 1 with at least some probability, $1 - n^{-3}/2$ (say).

Theorem 2. Set a matrix $U \equiv p^{-1}\|P_T P_\Omega P_T - pP_T\|$ and $\mathbb{E}\,U \leq 1$, one can get

$$\mathbb{p}\left(|U - \mathbb{E}\,U| > \delta\sqrt{\frac{\mu_0 nk\log n}{w}}\right) < 3\exp(-\varphi_0' \min\{\delta^2 \log n, \delta\sqrt{\frac{w\log n}{\mu_0 nk}}\}) \tag{3.23}$$

Candès and Recht (2008) establish that

$$U \leq C'\sqrt{\frac{\mu_0 nk\log n}{w}} + \frac{1}{\sqrt{\varphi_0'}}\sqrt{\frac{\mu_0 nk\beta\log n}{w}} \tag{3.24}$$

with probability at least $1 - 3n^{-\beta}$ when assuming $\delta = \sqrt{\beta/\varphi_0'}$ and

$w > (\beta/\varphi_0')\mu_0 nk\log n$. If w is approaching infinity, equation (3.22) can be expressed as:

$$\frac{p}{2}\|P_T(\boldsymbol{Q})\|_F \leq \|P_T P_\Omega P_T(\boldsymbol{Q})\|_F \leq \frac{3p}{2}\|P_T(\boldsymbol{Q})\|_F \tag{3.25}$$

for large $\boldsymbol{Q}$ with high probability. Equation (3.25) sets the lower and upper bound for the linear operator mapping $\boldsymbol{Q}$ from T to $\mathbb{R}^{n_1 \times n_2}$. As the potential solution matrix $\boldsymbol{Q}$ is large and low-rank matrix, there should be only one matrix in $\mathbb{R}^{n_1 \times n_2}$ when the Frobenius norm in a small range. Combining the duality and injectivity properties, we can conclude such matrix $\boldsymbol{Q}$ is unique in our case.

3.2 Convergence Analysis

The previous section shows the important work established by Candès and Recht (2008) and Candès and Tao (2009): a large matrix with low rank and a large proportion of missing values can be exactly recovered based on a small number of observed elements if the number of the observed entries is greater than the minimal sampling requirement. By exploring the properties of sampling operators and projections on the real number set and the location set, the solution matrix to this convex optimization problem is unique. Until now, we only knew there should exist a recovered and completed matrix corresponding to the ARMS data matrix $\boldsymbol{W}$ at some probability in theory. But how to find such a converged solution matrix in practice is still an impediment to overcome. We need to derive a mathematical solution to this convex optimization problem step by step. Fortunately, the study from Cai, Candès and Shen (2008) shows an efficient algorithm to find a numerical solution addressing this problem.

Recalling equation (3.1), I return to that equation as the basic idea to solve the missing data problem. However, equation (3.1) contains the tightest constraint because it is subjected to a one-element to one-element condition. Cai et al. (2008) develop a singular value thresholding method to solve this rank minimization problem based on iterative Newton steps. They relax the constraint and propose instead to solve

$$\begin{aligned} &\text{minimize: rank } (\mathbf{Q}) \\ &\text{subject to } A(\mathbf{Q}) = T \end{aligned} \tag{3.26}$$

where A is linear operator and $T \in \mathbb{R}^{n_1 \times n_2}$. The constraint in equation (3.26) is an obviously relaxed one compared to equation (3.1). The constraint evolves from one-element to one-element condition to a linear equation system. T can be a vector of values controlling the upper and lower bound in $\mathbb{R}^{n_1 \times n_2}$. The algorithm of Cai et al. (2008) can be expressed as:

$$\begin{cases} \boldsymbol{Q}^i = shrink\left(\boldsymbol{K}^{i-1}, \gamma\right) \\ \boldsymbol{L}^i = \boldsymbol{L}^{i-1} + \varphi_i P_\Omega(\boldsymbol{W} - \boldsymbol{Q}^{\mathrm{i}}) \end{cases} \tag{3.27}$$

In equation (3.27), $\boldsymbol{L}^{i-1}$ is the intermediate solution matrix at each level γ and $\boldsymbol{L}^{i-1}$ starts from $\boldsymbol{L}^0 = 0$. In addition, $shrink$ denotes the nonlinear function under singular value thresholding framework, γ indicates the corresponding level parameter and $\{\varphi_i\}$ denotes the sequence of step sizes. A nonlinear shrink function will pull down the largest eigenvalues and pull up the smallest eigenvalues until they converge to the grand mean of all sample eigenvalues. With the increase of level γ, $\boldsymbol{Q}^i$ converges to the stopping criteria which can minimize equation (1). A new operator is introduced in soft-thresholding algorithm (Cai et al., 2008):

$$D_\gamma(\boldsymbol{Q}) = \boldsymbol{V} D_\gamma(\boldsymbol{\Sigma}) \boldsymbol{U}^* \tag{3.28}$$

where $D_\gamma(\boldsymbol{\Sigma}) = diag\{(\sigma_j - \gamma)_+\}$, $\Sigma = \mathrm{diag}\{\sigma_j\}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices decomposed from $\boldsymbol{Q}$ as defined in previous section. Since we start from $\boldsymbol{L}^0 = 0$ in the Singular Value Thresholding (SVT) algorithm, we get new matrices $\boldsymbol{L}^\gamma$ and $\boldsymbol{Q}^\gamma$ at every level γ. Thus, σ_j, U and V change at different levels. The term $(\sigma_j - \gamma)_+$ indicates we choose the maximum value from $(0, \sigma_j - \gamma)$. If most of singular values in $\boldsymbol{Q}$ are less than the level γ, the soft-thresholding operator D_γ decays rapidly. Considering that the factors in D_γ are calculated based on $\boldsymbol{Q}$, the decay speed of D_γ reflects how far the solution matrix $\boldsymbol{Q}$ is away from stopping criteria. It is intuitive that the matrix closest to stopping criteria can be regarded as the best recovered matrix. Cai et al. (2008) propose the theorem to present such a proximity property of the soft-thresholding operator (proof for Theorem 3 is in appendix):

Theorem 3. For level γ, the singular value shrinkage operator is defined as:

$$D_\gamma = \mathrm{argmin}\{\tfrac{1}{2}\|\boldsymbol{Q} - \boldsymbol{L}\|_F^2 + \gamma\|\boldsymbol{Q}\|_*\}$$

Theorem 3 shows that the singular value shrinkage operator of solution matrix $\boldsymbol{Q}$ is an approximation of the nuclear norm of $\boldsymbol{Q}$ because the level γ is controlled during the calculation. Another reason is that the Frobenius norm is connected closely to the nuclear norm when evaluated for a large sparse matrix based on the representation in Xi Peng et al., 2015.

Substituting the singular value shrinkage operator D_γ in equation (27), we present the inductive algorithm as:

$$\begin{cases} \boldsymbol{Q}^i = D_\gamma\left(\boldsymbol{L}^{i-1}\right) \\ \boldsymbol{L}^i = \boldsymbol{L}^{i-1} + \varphi_i P_\Omega\left(\boldsymbol{W} - \boldsymbol{Q}^{\mathrm{i}}\right). \end{cases} \tag{3.29}$$

To apply a shrinkage operator in an iterative process, Cai et al. (2008) rewrite equation (3.2) based on the relationship between Frobenius and nuclear norms as follows:

$$\text{minimize: } \frac{1}{2}\|\boldsymbol{Q}\|_F^2 + \gamma\|\boldsymbol{Q}\|_* \tag{3.30}$$

$$\text{subject to } \mathrm{P}_\Omega(\boldsymbol{Q}) = \mathrm{P}_\Omega(\boldsymbol{W})$$

Since we are interested in recovering a large matrix such as the ARMS data matrix, we need to produce as long a sequence of singular values decomposed from true completed data matrix as possible. By setting a large level γ, we can ensure that the solution matrix $\boldsymbol{Q}$ converges to the true matrix. Another advantage to specifying equation (3.30) as an objective function is taking the sparsity of matrix into account. We assume we can observe as many elements in location set Ω as possible. This means $\boldsymbol{Q}^i$ should vanish outside Ω. By equation (3.30), we can check the sparsity of $\boldsymbol{Q}$ and ensure the elements lie outside Ω at every level.

Cai et al. (2008) propose a Lagrangian form of the problem statement using the basic idea of Uzawa's iteration algorithm. They design a function defined by different levels as: $f_\gamma(\boldsymbol{Q}) = \gamma\|\boldsymbol{Q}\|_* + \frac{1}{2}\|\boldsymbol{Q}\|_F^2$, which is the same as equation (3.30). The Lagrangian objective function for this problem is defined as:

$$\mathcal{L}(\boldsymbol{Q}, \boldsymbol{L}) = f_\gamma(\boldsymbol{Q}) + \langle \boldsymbol{L}, \mathrm{P}_\Omega(\boldsymbol{W} - \boldsymbol{Q}) \rangle \quad (3.31)$$

Uzawa iteration aims to solve saddle point problems in the context of concave optimizations, ensuring you find a global optimum. When we apply Uzawa iteration to equation (31), $\mathrm{P}_\Omega(\boldsymbol{W} - \boldsymbol{Q})$ and $\boldsymbol{K}$ are updated at the same time. If Uzawa iteration is applied to solve the linear system equation $A\boldsymbol{x} = b$, where $\boldsymbol{x}$ is a vector or matrix, the iteration process would be terminated when the norm of $\boldsymbol{x}$ is sufficiently small. With the properties of Uzawa iteration, we can finally get a primal-dual optimal vector pair $(\boldsymbol{Q}^*, \boldsymbol{L}^*)$ obeying:

$$\mathcal{L}(\boldsymbol{Q}^*, \boldsymbol{L}^*) = sup\ inf\ \mathcal{L}(\boldsymbol{Q}, \boldsymbol{L}) = inf\ sup\ \mathcal{L}(\boldsymbol{L}, \boldsymbol{Q})$$

Cai et al. (2008) define a dual function when $\boldsymbol{Q}$ approaches infinity in Lagrangian function. The function is $\boldsymbol{z_0}(\boldsymbol{L})=inf_{\boldsymbol{Q}}\mathcal{L}(\boldsymbol{L}, \boldsymbol{Q})$. Lagrangian function becomes:

$$\partial_{\boldsymbol{L}}\mathcal{L}\left(\boldsymbol{L}, \widetilde{\boldsymbol{Q}}\right) = \partial_{\boldsymbol{L}}\boldsymbol{z_0}(\boldsymbol{L}) = \mathrm{P}_\Omega(\boldsymbol{W} - \widetilde{\boldsymbol{Q}}) \quad (3.32)$$

In equation (3.32), $\widetilde{\boldsymbol{Q}}$ is updated through Uzawa iteration process as the minimizer to equation (3.30). Before applying the Uzawa iteration process to SVT convergence analysis, we should notice that

$$\arg\min f_\gamma(\boldsymbol{Q}) + \langle \boldsymbol{L}, \mathrm{P}_\Omega(\boldsymbol{W} - \boldsymbol{Q}) \rangle = arg\ min\ \gamma\|\boldsymbol{Q}\|_* + \frac{1}{2}\|\boldsymbol{Q} - \mathrm{P}_\Omega\boldsymbol{L}\|_F^2 \quad (3.33)$$

Equation (33) builds up the connection between our initial problem statement and the Lagrangian form. Then, the objective function becomes:

$$\begin{aligned} &\text{minimize: } f_\gamma(\boldsymbol{Q}) \quad (3.34)\\ &\text{subject to } A(\mathbf{Q}) = T \end{aligned}$$

where the adjoint matrix of A is A^*. The Lagrangian expression in equation (34) is $\mathcal{L}(\boldsymbol{l}, \boldsymbol{Q}) = f_\gamma(\boldsymbol{Q}) + \langle \boldsymbol{l}, T - A(\mathbf{Q}) \rangle$. We can start an Uzawa iteration process from $\boldsymbol{l}^0 = \mathbf{0}$. Since $\boldsymbol{L}^i = \boldsymbol{L}^{i-1} + \varphi_i P_\Omega(\boldsymbol{W} - \boldsymbol{Q}^{\mathrm{i}})$ and $A^*A = \mathrm{P}_\Omega$, $\boldsymbol{L}^i = A^*(\boldsymbol{l}^i)$, we can calculate equation (29) again by substituting $\boldsymbol{L}^i$ and P_Ω.

In the previous section, we show the Lagrangian specification of the objective function. In fact, the constraint in equation (29) can also be specified in another form. Cai et al. (2008) propose the general convex constraint on minimizing the nuclear norm of solution matrix $\boldsymbol{Q}$. Assume there is a convex set C given by

$$C = \{f_i(\boldsymbol{Q}) \leq 0, \forall i = 1, \ldots, n\} \tag{3.35}$$

where f_i denotes the *i*th convex function. We should be clear that f_i in equation (3.35) is different from f_γ in the objective function. In equation (3.35), f is a convex function with no detailed specification and we use *i* to denote each convex function at each update step within the Uzawa iteration. However, in our objective function, we are pretty sure that function f_γ is defined as $\gamma\|\boldsymbol{Q}\|_* + \frac{1}{2}\|\boldsymbol{Q}\|_F^2$ at each level γ. We set up a set containing all the convex functions f_i as: $\mathcal{F}(\boldsymbol{Q}) = (f_1(\boldsymbol{Q}), \ldots, f_i(\boldsymbol{Q}))$. The Lagrangian form of our problem is $\mathcal{L}(\boldsymbol{l}, \boldsymbol{Q}) = f_\gamma(\boldsymbol{Q}) + \langle \boldsymbol{k}, \mathcal{F}(\mathbf{Q}) \rangle$. We are sure that all $\boldsymbol{l}$ are greater than 0 by starting from $\boldsymbol{l}^0 = \mathbf{0}$ if we specify the objective function in Lagrangian form. This fact will lead to:

$$\begin{cases} \boldsymbol{Q}^i = arg\ min\ \{f_\gamma(\boldsymbol{Q}) + \langle \boldsymbol{l}^{i-1}, \mathcal{F}(\boldsymbol{Q}) \rangle\} \\ \boldsymbol{l}^i = [\boldsymbol{l}^{i-1} + \varphi_i \mathcal{F}(\boldsymbol{Q}^i)]_+ \end{cases} \tag{3.36}$$

In equation (36), $arg\ min\ \{f_\gamma(\boldsymbol{Q}) + \langle \boldsymbol{l}^{i-1}, \mathcal{F}(\boldsymbol{Q}) \rangle\} = D_\gamma\left(A^*(\boldsymbol{l}^{i-1})\right)$ and $\mathcal{F}(\boldsymbol{Q}^i) = (T - A(\mathbf{Q}))$. The symbol $[a]_+$ denotes the maximum value in (a,0). By equation (3.36), we can get an intuitive understanding about the convergence property. At level γ, we get the minimum value of $\boldsymbol{l}^{i-1}$ and potential solution matrix $\boldsymbol{Q}$. Under the convex set $\mathcal{F}$, we have the maximum value of $\boldsymbol{l}^i$ for level $\gamma + 1$. In next level, we update $\boldsymbol{Q}$. This process is similar to process of searching for saddle points in a convex set. In the first step, we find candidate points closest to the local minimum or maximum values under certain constraint. By relaxing the

constraint and testing the distance to the new local extremum values, we get updated values of saddle points. Then, we repeat such a search process until we can't relax the constraint any more.

Cai et al. (2008) propose two theorems and two lemmas to make the convergence property more general and transparent. Lemma 7 and Theorem 4 establish a simple convergence proof for equation (3.27). Lemma 8 and Theorem 5 establish convergence proof for equation (3.34) when the solution matrix is the one in equation (3.36). (All the proof work is done by Cai et al., (2008) in the appendix)

Lemma 7. If $\boldsymbol{X} \in \partial f_{\gamma}(\boldsymbol{Q})$ and $\boldsymbol{X}' \in \partial f_{\gamma}(\boldsymbol{Q}')$. Then $\langle \boldsymbol{X} - \boldsymbol{X}', \boldsymbol{Q} - \boldsymbol{Q}' \rangle \geq \|\boldsymbol{Q} - \boldsymbol{Q}'\|_F^2$.

Lemma 8. Assume $(\boldsymbol{Q}^*, \boldsymbol{l}^*)$ is optimal pair derived from equation (34), $\boldsymbol{l}^*$ obeys $\boldsymbol{l}^* = [\boldsymbol{l}^* + \varphi \mathcal{F}(\boldsymbol{Q}^*)]_+$ when $\varphi > 0$.

Theorem 4. If the step size follows $0 < \inf \varphi_i < sup\ \varphi_i < 2$. Then the solution matrix $\boldsymbol{Q}$ derived from equation (3.27) is converged to the solution in equation (3.30).

Theorem 5. Let $\mathcal{L}(\mathcal{F})$ be a Lipschitz constant and $0 < \inf \varphi_i < sup\ \varphi_i < 2/\|\mathcal{L}(\mathcal{F})\|^2$, then the solution matrix derived from equation (3.36) converges to the solution in equation (3.34).

Based on the previous analysis and proof, we can conclude that the solution matrix $\boldsymbol{Q}$ to recover the target matrix $\boldsymbol{W}$ is the unique and converged solution when we apply the SVD algorithm and set a singular value threshold on solution matrix $\boldsymbol{Q}$.

CHAPTER 4

DATA AND EXPERIMENTAL DESIGN

This study uses three datasets. The first one is a large fake dataset. This dataset is created by ourselves. Candès and Recht (2008) prove the matrix completion approach is valid under a normal distribution theoretically. Thus, we create a completed dataset where each column is generated by standard normal distribution. We assign missing values randomly and apply the chosen matrix completion approach to get a recovered matrix. Then, we compared this recovered matrix to completed matrix and estimate the kernel distribution of each column. Because we want to estimate the distribution of every main commodity, we can validate the estimated density is reliable after matrix completion approach when we are clear about the data generation process.

The second dataset is Simulated Farm Data. The representative farm data come from the Agricultural and Food Policy Center at Texas A&M University. The dataset is based on actual farm data from 2016 and 2017 for a selection of that center's representative farms that grow corn, soybeans, and/or wheat. Simulations were then used to generate a large set of data based on the likely outcomes from those farms under a large number of random weather, pest, and market conditions. The end result is 5,000 simulated observations that should be very characteristic of large, commercial farms. We name this dataset as "Simulated Farm Data" for simplicity. For every farmer in this dataset, we "know" the "true" values for acreage, commodity-specific costs and other necessary statistics of every commodity. We can calculate total cost, revenue and net farm income of every commodity. We can then replace a percentage of true values by missing

values in this dataset and use those low-rank matrices to test our matrix completion approach. Because we know all the elements in this data matrix exactly, we can check how far the recovered values deviate from the true values on data that should be very similar to true U.S. farm-level data such as the ARMS dataset.

The third dataset is built from the Agricultural Resource Management Survey (ARMS). The ARMS dataset includes all the financial information of farmers, including costs and revenues. However, while the revenue information is commodity-specific, input cost and usage variables in the ARMS dataset are reported only for the sum of all the commodities. Thus, in general, we don't know the cost and net farm return of specific commodities from the ARMS dataset. Some farmers may produce only one commodity. Their net farm income can be regarded as the net return of one commodity. But for other farmers who grow more than one commodity, the net returns for specific commodities are missing. The ARMS dataset is the target matrix to which we want to apply the matrix completion approach, thereby recovering the data on commodity-specific returns. After we recover the missing values in the ARMS data, we can estimate the distribution of net farm return for each commodity which would be very useful for policy analysis.

4.1 Fake Dataset

In this part, we construct a fake data matrix as the first step to develop our research. The fake data matrix is a 2000×200 matrix, denoted as $\boldsymbol{X}$. We generate all values in each column randomly from a standard normal distribution. Then we assign positions for missing values randomly and replace the true values with NAs. We get the incomplete matrix $\boldsymbol{X_{inc}}$ and try to recover it. We define the missing proportion, m_p, to indicate the proportion of missing values in fake data matrix. We increase missing proportion from 25% to 70% by 0.5% at a time.

When we assign the positions to missing values, the sequence of position indicators is generated randomly. This should avoid too many missing values in one column. Since every position indicator is generated independently, it is not probable that too many consecutive position indicators would be generated.

By applying the matrix completion approach, we get the recovered matrix, denoted as $\boldsymbol{X}_{rec}$. We already know the positions of missing elements in $\boldsymbol{X}$; thus, we can pick out the recovered values in $\boldsymbol{X}_{rec}$ based on the positions assigned to the missing values and denote them as $\boldsymbol{X}_{recm}$. Denote the original values at the missing value positions in original matrix $\boldsymbol{X}$ as $\boldsymbol{X}_m$. To check the distance from recovered matrix to original matrix, we define the recovery percentage error as $\boldsymbol{A} = |(\boldsymbol{X}_{recm}[i,j] - \boldsymbol{X}_m[i,j])/\boldsymbol{X}_m[i,j]|$, where *i* is row index and *j* is column index.

We can divide $\boldsymbol{A}$ into different groups and calculate the percentages of these groups based on the values of base lines we choose. We first set up the base line as $1e-4$. The values in $\boldsymbol{A}$ less than $1e-4$ are denoted as "T", which means they are truly below the base line. The values larger than $1e-4$ are denoted as "F", which means they are above the base line. We calculate the proportion of "T" and "F" in $\boldsymbol{A}$. We change the values of base lines to be $1e-3$, $1e-2$ and $1e-1$ respectively. For each cutoff value, we calculate the percentages of "T" and "F" in $\boldsymbol{A}$. We get different percentages of "T" and "F" for different base lines at one missing proportion. We repeat such process for different missing proportion in matrix $\boldsymbol{X}$ and calculate the corresponding percentages of "T" and "F".

4.2 Simulated Farm Data

There are 5000 observations in Simulated Farm Data. There are three commodities in this dataset: corn, wheat and soybean. If farmers don't grow one commodity, all the statistics about this commodity are denoted as zeros. All information is completed except total production cost

for each commodity. The cost variables reported on a per acre basis are: seed, fertilizer, fuel, chemicals, irrigation and so on. The information on these cost variables for all observations are completed. But the fixed cost for each commodity grown in one farm is missing. We only know the total fixed cost for the whole farm basis instead of commodity. We need to compute the net return of each commodity by the following steps:

1. We denote the total fixed cost as TFC_i and total variable cost as TVC_i, where *i* denotes *i*th farmer. The total cost $TC_i = TFC_i + TVC_i$. As we know exact acreages of corn, soybean and wheat for each farmer, we can define the acreage share as:

$$Share_{n,i} = Acre_{n,i}/(Corn\ Acre_i + \ Soybean\ Acre_i + Wheat\ Acre_i) \quad (4.1)$$

where *n* = {corn, soybean, wheat}.

2. For the fixed cost of per acre for each commodity as:

$$Cost_{n,i} = (Seed_{n,i} + Fert_{n,i} + Chem_{n,i} + Fuel_{n,i} + Water_{n,i} + Other_{n,i}) \quad (4.2)$$

where *n* and *i* are the same in equation (1).

3. The total cost for each commodity is defined as:

$$TC_{n,i} == TFC_{n,i} + TVC_{n,i} = Cost_{n,i} \times Acre_{n,i} + TFC_i \times Share_{n,i} \quad (4.3)$$

4. For each commodity, we know the unit price, receipt, yield per acre, total net farm income and so on. Thus, the net return of each commodity in every farm can be calculated as:

$$Return_{n,i} = Price_{n,i} \times Acre_{n,i} \times Yield_{n,i} - TC_{n,i} \quad (4.4)$$

where $Price$ denotes the unit price for each commodity, $Acre$ denotes the total acre of commodity grew by each farmer, $Yield$ denotes the production per acre of commodity and *n* and *i* are the same as equation (3.1).

After we sum the net return of three commodity for each farm, we confirm that this summation is equal to the farmer's net farm. Our finding supports that the approach to calculate

the net return of each commodity is reliable. The reason why we use Simulated Farm Data is that this dataset is complete and net return of each commodity can be computed exactly. If the commodity is not planted, there are zeros not missing values. This representative farm dataset plays an important role to validate that our matrix completion approach can recover the missing values in real dataset accurately. The validation process on the representative farm data is quite close to the test on the fake data matrix, but adds assurance that the algorithm works well on data that has characteristics such as we expect to find in the ARMS data. During the calculation of commodity-specific per acre net return in Simulated Farm Data, we observe some unexpectedly large and small values. To eliminate undue influence of such outliers, we truncate the net returns for the three commodities and get a subset of original Simulated Farm Data with all the information known but extreme tail observations removed. We denote this subset of Simulated Farm Data as $\boldsymbol{T}$. As we get start with a complete dataset, we again assign positions for missing values randomly and insert NAs into the data matrix. We denote this incomplete matrix as $\boldsymbol{T}_m$. In this matrix, we increase the missing proportion from 20% to 90% by 0.5% increments. Our initial goal is to find the distribution of net return for each commodity. Thus, missing value positions are only assigned among the columns of net returns for each of the three commodities. After we recover the matrix, we get the recovered matrix $\boldsymbol{T}_r$. We focus on differences between the recovered values of net return and the original values of net return. We extract recovered values in the net return columns of corn, soybean and wheat in the recovered matrix and denote this subset as $\boldsymbol{T}_{rec}$. We extract the original values in net return columns of corn, soybean and wheat in $\boldsymbol{T}$ and denote this subset as $\boldsymbol{T}_c$. We use two measurements, Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE), to check the recovery effect after recovering fake data matrix. We define Mean Absolute Percentage Error as $\boldsymbol{MAPE} =$

$\frac{100\%}{n}\sum_{i=1}^{n}|\boldsymbol{T}_{rec}[i,j]-\boldsymbol{T}_{c}[i,j]/\boldsymbol{T}_{c}[i,j]|$ and Mean Absolute Error (MAE) as $\boldsymbol{MAE}=\frac{\sum_{i=1}^{n}|\boldsymbol{T}_{rec}[i,j]-\boldsymbol{T}_{c}[i,j]|}{n}$. MAPE measures deviation between recovered values and true values in percentage while MAE just measures the difference between the recovered values and true value to indicate the recovery accuracy in terms of dollars per acre. We also try to fit the different distributions to the net return of the three commodities to check whether the commodity net return distribute as specific distribution.

4.3 ARMS Dataset

ARMS is an annual survey covering 19623 observations in the 2014 edition used here. The respondents include small-scale farmers, large-scale farmers and large corporate farms. All agricultural commodities are included in this dataset: corn, soybean, wheat, hog, cattle, cotton, hay, barley, oats and so on. The main statistics for these commodities are production quantity, sales, income, acres planted and harvested, cost/expense, and so on. In the ARMS dataset, the net farm income, the cash income under contract and expenses are collected as the sum for all commodities. Take the total expense as an example to illustrate the challenge in ARMS dataset. For one farm household, we already know the total expense. When we consider the cost of growing corn, we only know the quantity of corn produced and acres used; we don't know the fertilizer cost, fuel cost, insurance cost and other statistics specifically related to corn production. The same dilemma happens to the net farm return of each commodity in the ARMS dataset. If we want to estimate distribution of commodity-specific net return, we need to separate the net return for each commodity.

To focus on the subset of (roughly) commercial farmers, we create a smaller dataset denoted as $\boldsymbol{W}_0$. In dataset $\boldsymbol{W}_0$, all the farmers farm over 100 acres. The total number of farmers in $\boldsymbol{W}_0$ is 14405. $\boldsymbol{W}_0$ includes net farm income, operational expenses and production of each respondent

from the ARMS dataset. We also extract production levels and acres related to the commodities from the ARMS dataset into $\boldsymbol{W}_0$. Because there are so many commodities in $\boldsymbol{W}_0$, we only focus six main commodities: corn, soybean, wheat, hog, cattle and cotton. As explained in the methodology part, there should not be too many zeros in the net return columns of these six commodities. To investigate the distributions of the net return of the main commodities, we extract the farmers who operate only one of the six commodities.

Take farmers who only grow corn as examples. If there is a subgroup of farmers who only grow corn, their net farm income can be regarded as net return of growing corn. We can create a new column to represent the net return of corn. In this column, the net returns of corn for farmers who only grow corn are their corresponding net farm incomes and the net returns of corn for other farmers are regarded as missing values. We use NA to denote missing value in our dataset. By similar process of creating net return column for corn, we can create another five new columns to represent net returns of soybean, wheat, hog, cattle and cotton. We combine these six new columns with $\boldsymbol{W}_0$ to make up our target matrix $\boldsymbol{W}$. Based on Table 4.1, we can conclude that the number of farmers who operate only one commodity is very small which makes matrix $\boldsymbol{W}$ is a sparse matrix.

CHAPTER 5

RESULTS

5.1 Fake Dataset

Figure 5.1 presents the intuitive range of values in $\boldsymbol{A}$ with different missing proportions in the matrix. The deviation rate in Figure 5.1 shows how far away the recovered values are from the true values. The densest part of these dots is just above zero. From Figure 5.1, we have a rough idea that the most of recovered values are close to the true values when we test the matrix completion algorithm on the fake data matrix. We can obtain more details about recovery effect from Table 5.1. By defining "T" and "F" groups to measure recovery effect, we count the percentage ofthat recovered values in group "T" as being "close enough" to the true values while recovered values in group "F" are "far away" from the true values. For example, if the missing proportion is 0.205, the percentage of group "T" is around 0.525182 and the percentage of group "F" is around 0.474818 when the when the value of base line is $1e-04$. When missing proportion is fixed as 0.205, the percentage of group "F" is around 0.474818 if the value of base line is $1e-04$. However, the percentage of group "F" is around 0.470695 if the value of base line is $1e-01$ with the same missing proportion. The results in Tables 5.1 show that when the value of base line is fixed, the percentage of group "T" is always larger than the percentage of group "F" with the increase of missing proportion. When the missing proportion is fixed, the percentage of group "F" is decreasing when the value of base line increase. When the missing proportion is 0.25 and the value of base line is $1e-04$, the percentage of group "T" is around

0.522950. If the missing proportion is 0.55 and the value of base line is $1e-04$, the percentage of group "T" is around 0.521245. For the different missing proportions, the percentage of group "T" is always around 0.52, larger than 0.5. Combining the above findings together, we can conclude the most of recovered values can be regarded as close enough to their true value. The recovery effect is good if we apply the matrix completion approach on our fake data matrix. There is an overall trend that the recovery effect would be worse when the missing proportion is very large. The reason is that we can get les information on the target matrix $\boldsymbol{X_{inc}}$ when $\boldsymbol{X_{inc}}$ is extremely sparse.

Our research goal is to investigate the distributions of net return of each commodity. When we use the fake dataset, such a research goal can be achieved by estimating probability density function of every column in recovered matrix $\boldsymbol{X_{rec}}$. Since every column in the fake dataset is generated randomly from a standard normal distribution, the density estimation process for each column is the same. We just take one column as an example. Figure 5.2 shows the result of kernel density estimation. It seems that values in that column are distributed as Gaussian distribution.

Next, we do the Anderson-Darling Test to verify the normality hypothesis. The Anderson-Darling (A-D) test is a goodness-of-fit test which is used for deciding whether a sample is drawn from a specified distribution or not, most commonly whether the sample data is drawn from the normal distribution (M.J. De Smith, 2014). The null hypothesis of the Anderson-Darling (A-D) test is that the sample is drawn from the normal distribution. Our test result shows that the p-value is 0.1897. It fails to reject the null hypothesis and we can conclude that the values are distributed as normal distribution. Since we generate the elements in the fake dataset from

standard normal distributions, the test result is reasonable. We believe that matrix completion approach can recover the missing values accurately when we use the fake data to validate it.

5.2 Simulated Farm Data

Table 5.2 presents the summary statistics on net returns of every commodity in Simulated Farm Data after we cut the tails. These net returns are measured on a per acre basis. The average net returns of corn, soybean and wheat are all larger than zero. These values seem to be consistent with our knowledge on daily operations of farmers.

Table 5.3 shows Mean Absolute Percentage Error (MAPE) on recovered commodity net return in Simulated Farm Data as the missing proportion is increasing. For example, Table 5.3 reports that MAPE is around 0.5812% when the missing fraction is 0.205. Since MAPE can provide us information about how far the recovered values differ from the true values, Table 5.3 suggests that the recovery accuracy is very high because most MAPE values are smaller than 10%. We believe the matrix completion approach is effective when we apply it to Simulated Farm Data. However, plot (a) in Figure 5.3 shows that there is no obvious trend of MAPE values when the missing proportion is increasing. In addition, there are some unexpectedly large MAPE values which are greater than 15% in Table 5.3. Some true values extracted from commodity net returns in the Simulated Farm Data are very close to zero. The divisions are quite large if the differences between recovered values and true values are divided by very small true values, which leads to unexpectedly large MAPE values. We also use Mean Absolute Error (MAE) to measure the recovery accuracy. The calculation of Mean Absolute Error is similar to calculation of Mean Absolute Percentage Error. But Mean Absolute Error avoids dividing a small value close to zero. Results in Table 5.4 show that MAE value is increasing as the missing proportion is increasing. Plot (b) in Figure 5.3 also shows the obviously upward tendency of MAE value with the

increasement of missing proportion. This is consistent with our assumption that the differences between recovered values and true values become larger when the missing proportion increases.

We try to fit different potential distributions to commodity-specific per acre net return in Simulated Farm Data. The test statistics and corresponding p-values are listed in Table 5.5. All the p-values are less than 0.05. Thus, there are no specific distributions which can fit the commodity-specific per acre net return in Simulated Farm Data perfectly. The irregular distributions of commodity-specific per acre net returns may be the reason why MAE values are not strictly increasing monotonically with the increasement of missing proportions. However, there is still an obviously upward trend of MAE values. Our findings based on MAPE values and MAE values suggest that matrix completion approach is reliable and stable when applying it to the (truncated) Simulated Farm Data. The success of recovering commodity-specific net returns in Simulated Farm Data is supportive enough to conclude that we can recover commodity-specific returns from the ARMS dataset precisely enough for use in policy analysis by the matrix completion approach.

5.3 Agricultural Resource Management Survey (ARMS) Data

Table 5.6 presents some basic statistics about the total net returns of the six main commodities from the ARMS dataset after recovering the missing values. For the average net return per acre of each commodity except that of hog, all the values are reasonable compared to the same statistics derived from Simulated Farm Data in Table 5.2. The average hog net return per animal and average hog net return for each farmer are negative. However, the maximum hog total net return for one farmer is quite large. The strong contrast between the average and maximum of

hog net return can be explained by the contract between large corporations and farmers[1]. The large hog corporations such as Tyson Food are counted in the ARMS dataset. That is why the maximum value of net returns can be very large. This negative effect on raising hogs caused by the farmers' small scales should be considered when U.S. government makes agricultural policy.

Figure 5.4 to Figure 5.9 present histograms for total net returns of each commodity before and after missing-value recovery. Plot (a) in each figure show the overall trends of commodity-specific total net returns before imputation. These plots indicate the available data on commodity net returns are very sparse compared to the total number of observations in the ARMS dataset. Plot (b) in each figure presents the histogram of total net return on specific commodity after missing data recovery. In Figure 5.4 to Figure 5.9, the range of every commodity's total net returns in plot (b) is much larger than the corresponding range in plot (a). We can also conclude that the overall trend of total net returns of each commodity in plot (b) is different from that in plot (a) from Figure 5.4 to Figure 5.9. Table 5.7 presents the results of two sample K-S test on every commodity's total net return before and after missing data recovery. For the total net return of each commodity, the p-values are all less than 0.05 and we reject the null hypothesis that the two distributions come from same distribution. The results in Table 5.7 indicate that the distribution of every commodity's total net return before missing data recovery is different from the distribution of total net return of each commodity after missing data recovery. Combined with the basic statistics in Table 5.6, our findings suggest that the recovered values contain more variation than observed values. The generation process of recovered values doesn't heavily rely on the available data. The missing-value recovery process in our case is not simply choosing

[1] The large hog corporations sign contracts with small-scale farmers. They may ask farmers to turn in hogs five times a year. However, the conditions in small farms might be too hot or too cold to raise hogs. The number of hogs turned in by the small-scale farmers is less than the amount listed in the contract. The famers make less revenue than their expectation while their total cost may increase. Thus, the average hog net return per acre and average hog net return for each farmer are negative.

repeated values from the observed data randomly. Comparison between plot (a) and (b) in every figure and the results in Table 5.7 strongly indicate that the imputation mechanism in matrix completion approach is totally different from that in either hot deck imputation or a Bayesian approach.

Informed by the application of matrix completion approach to the Simulated Farm Data, we also focus on the central part of commodity-specific per acre net return or per animal net return for the ARMS dataset. After we get all the recovered values for net return per acre or net return per animal, we focus on the center part of these recovered values. Table 5.8 presents summary statistics on the central part of the distributions of commodity-specific per acre net return or per animal net return after missing data recovery. The average per animal net return for raising cattle is the largest net return. All the means and variances of commodity-specific per acre net return or per animal net return are reasonable and consistent with our knowledge. As recovered values for commodity-specific per acre net return or per animal net return are dense, we estimate various percentiles of these recovered values. Table 5.9 shows percentile estimation on the central part of commodity-specific per acre net return or per animal net return after missing data recovery. Table 5.9 give us a good idea about the level of individual farmer's commodity-specific per acre net return at different percentiles of the distribution. For example, we can infer that one farmer gets much a higher per acre net return of corn than almost all farmers if the corn per acre net return for this individual famer is 90 dollars. If we are only interested in corn per acre net return, around 50% farmers get net returns smaller than or equal to $71.53 per acre.

We also want to investigate the distribution of commodity total net returns, we should employ more accurate measurement techniques to estimate the kernel density. Figure 5.10 to Figure 5.15 present plots fitting different distributions to total net returns of every commodity. Figure 5.10

indicates that normal distribution and log normal distribution are closer to the estimated empirical distribution corn net returns. Based on Figure 5.11, it is difficult to find a suitable distribution to fit the estimated wheat total net returns. We may use other approaches to figure out the best fitted distribution. Figure 5.12 shows that the normal distribution is the best choice to fit the soybean total net returns. From Figure 5.13, there is no strong evidence to select a suitable fitted distribution of hog total net returns. Figure 5.14 shows that either the log normal or gamma distributions could be potential kernel density distributions of cattle net returns. We can't derive any information about the fitted distribution of cotton net returns from Figure 5.15. We need to develop further analysis on the density estimation of cotton net returns. From Figure 5.10 to Figure 5.15, we can get an intuitive understanding about the density estimation of commodity net returns. We notice that there is no distribution that fits the recovered values of commodity total net returns perfectly. We thus turn to the Kullback-Leibler divergence measure to find the theoretical distribution with minimum distance to the estimated empirical distribution of each commodity's total net returns.

The Kullback-Leibler (KL) divergence can measure how one probability distribution diverges from the expected null distribution over the same variable *x*. After we get all the recovered values of commodity net returns, we can get estimated parameters for different distributions by fitting these distributions to the recovered data. Then we use the KL divergence measure to quantify the divergence of the estimated distributions from the theoretical distributions. Although KL divergence is not a metric measure, it can be a measurement of distance between "true" distribution and the actual distribution derived from the data. Table 5.10 provides KL divergence measuring the total net returns of different commodities. It suggests the normal distribution is the closet theoretical distribution to fit the total net returns distributions of five of the six

commodities studied: corn, soybeans, hogs, cattle, and cotton. When considering wheat total net returns, it is better to choose the log normal distribution to depict the overall distribution of wheat net returns. When we consider the fitted distributions of net returns for hogs and cattle, the gamma distributions might be worth exploring, even though it does not fit as well as the normal.

CHAPTER 6

CONCLUSION AND DISSCUSSION

Our research investigates the efficiency of a matrix completion approach to recover missing values in a practical dataset, applies this approach to recover the missing values in the ARMS dataset and estimates the kernel densities for the net returns of six major commodities. We explain why Bayesian approaches and Hot Deck Imputation are not suitable to apply to complete the missing commodity-specific net returns in the ARMS dataset. The disadvantage of Bayesian methods to imputing the missing value is that the accuracy of recovery effect heavily depends on the choices of model specification and prior distribution. The drawback of Hot Deck Imputation is that such approach is simply replace the missing values by the median or mean, which is not feasible to be applied in ARMS dataset where the missing proportion of missing values quite large. Further, both approaches would likely struggle given that the few observations we have on commodity-specific returns—from single commodity farms—are not representative of farms that produce multiple commodities.

Theoretically, we explore the framework of matrix completion approaches to show that it is very possible to recover a large and sparse matrix exactly by searching for a converged solution as long as the number of observed entries meets the minimal sampling requirement. We first apply the matrix completion approach to a fake dataset where every element is generated from a standard normal distribution and missing values are assigned randomly. Around 52% of the recovered values can be regarded as nearly the same as the true values in the initial completed

dataset. The recovery effect is still good when the missing proportion is very large. Results from Anderson-Darling tests are consistent with the fake data generation process. Our findings support that matrix completion approach is efficient at recovering missing values for a large and low-rank fake data matrix. It is reliable for us to do density estimation analysis on the columns of the recovered data matrix after we impute all the missing values.

Our study suggests that the matrix completion approach is also successful when applied to a simulated farm dataset generated from a set of representative farms. By computing per acre net returns of commodities in this dataset, we construct a completed and practical dataset whose structure is nearly the same as that of the ARMS dataset. The measurements Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) we employ to measure the recovery accuracy are satisfactory when we remove the tails of the distribution. All the above findings are strong evidence that the matrix completion approach is sufficiently effective at recovering missing data when it's applied to a dataset based on actual farm financial data collected through field survey.

Having confirmed the reliability of the method on two different datasets, we move on to the research goal: generation of commodity-specific net returns from the ARMS dataset that never collected the necessary data to compute those variables. After we impute the missing values in the ARMS dataset, we find that the average total net return on hogs is negative while total the net returns of corn, wheat, soybean, cattle and cotton are positive. We focus on the central part of the distribution of commodity-specific per acre net returns or commodity-specific per animal net returns. We also calculate different percentiles for commodity-specific per acre net return or commodity-specific per animal net return. We can estimate the percent level of commodity-specific per acre net return or per animal net return for an individual farmer. We find that to fit a

parametric distribution to the empirical distribution of commodity-specific total net returns, the normal distribution is a good choice for most commodities. For each of these six commodities, there is no specific distribution which can depict the overall trend of its total net return perfectly. We use Kullback-Leibler divergence to find the "closest" theoretical distribution to the recovered values of commodity total net return. Based on this metric, the normal distribution is a good choice for corn, soybeans, hogs, cattle, and cotton, while the log-normal distribution is the best candidate for the kernel density of wheat total net returns. The parameters of the fitted distributions differ from one commodity to another, however, even if they are mostly normally distributed.

Overall, our research provides an applicable approach to solve the shortcoming of missing commodity-specific returns in the ARMS dataset, allowing for more accurate analysis of agricultural policy in United States. The matrix completion approach can recover a large low-rank matrix accurately even when the missing proportion is large. Taking advantage of such techniques can allow us to extract much more information from existing government datasets. For example, knowing the percent of farms that are earning per acre net returns above some level would allow policy makers to determine the impact of changes in subsidy levels on not just aggregate or average farm income but also on individual farms. Knowing how a policy change would impact the percent of farms earning positive profits would be very valuable. The matrix completion approach applied here to ARMS data would allow such questions to be addressed.

Table 4.1 Summary Statistics for Farmers Who Operate Only One Commodity in ARMS Dataset

	Total observation	Mean acre	Mean income
Corn	103	403.4854	77409.28
wheat	148	780.8176	31322.19
soybean	157	559.8981	26366.29
hog	45	17652.64	1796722
cattle	1430	852.6958	104080.5
cotton	108	1155.972	68340

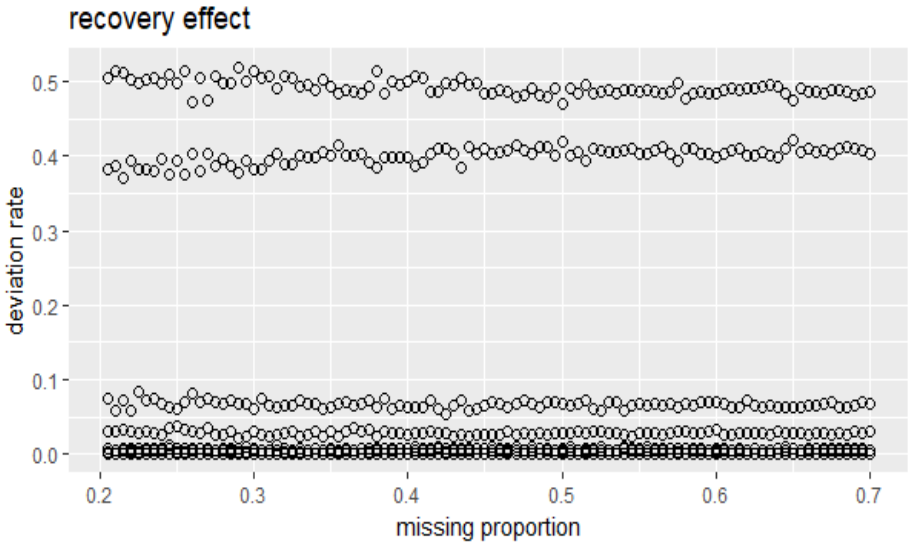


Figure 5.1 Intuitive Recovery Effects for Different Missing Proportions

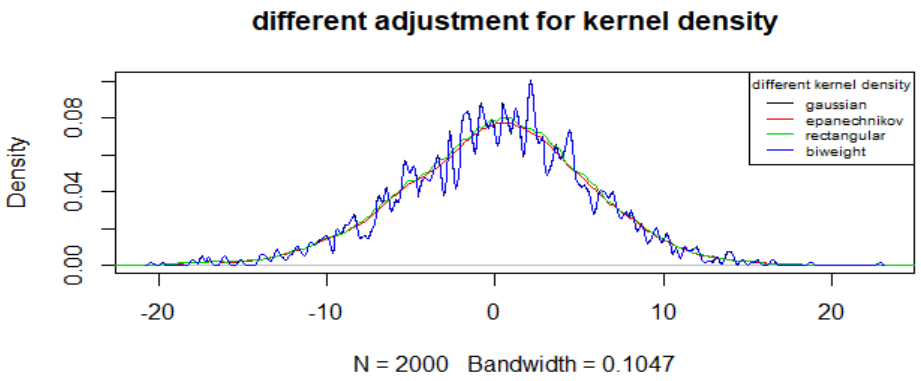
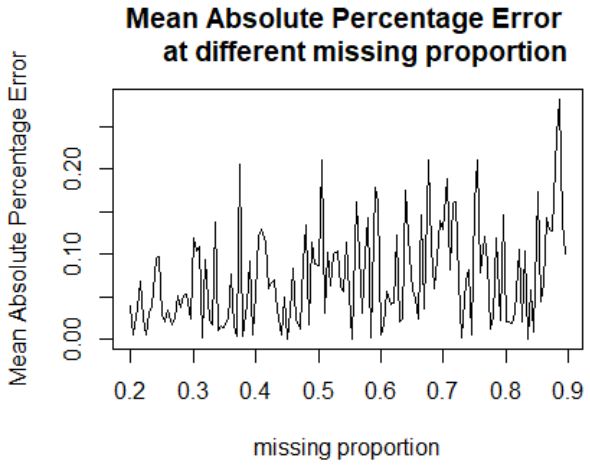


Figure 5.2 Density Estimation with Different Kernels

(a) MAPE on Commodity Net Return in Simulated Farm Data after Missing Data Recovery



(b) MAE on Commodity Net Return in Simulated Farm Data after Missing Data Recovery

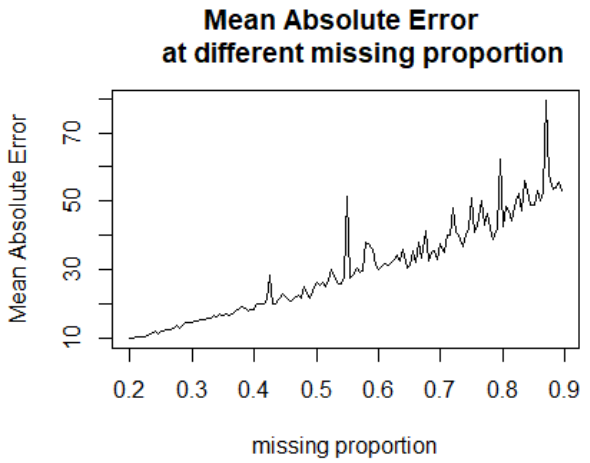
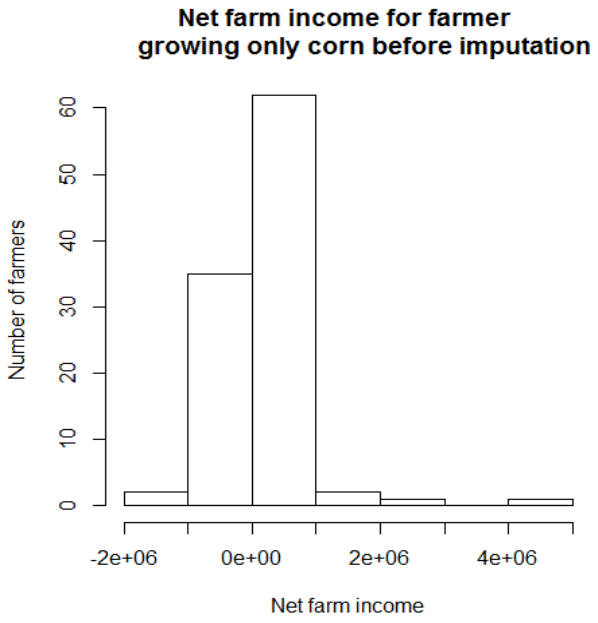


Figure 5.3 MAPE and MAE plots on Commodity Net Return in Simulated Farm Data after Missing Data Recovery at Different Missing Proportion

(a) Corn Total Net Return before Imputation



(b) Corn Total Net Return after Imputation

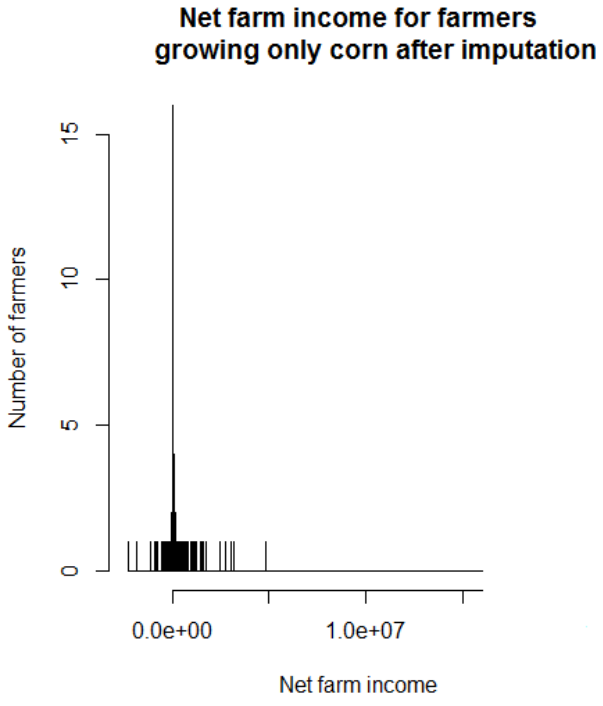
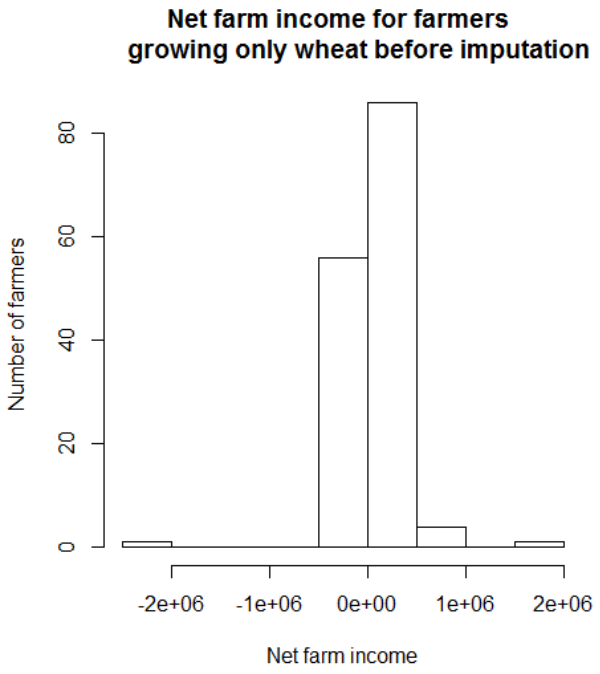


Figure 5.4 Corn Total Net Return Imputation

(a) Wheat Total Net Return before Imputation



(b) Wheat Total Net Return after Imputation

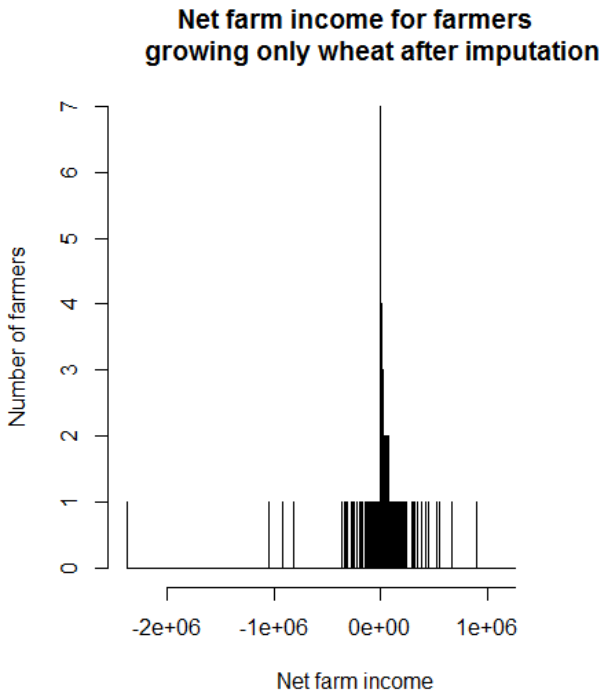
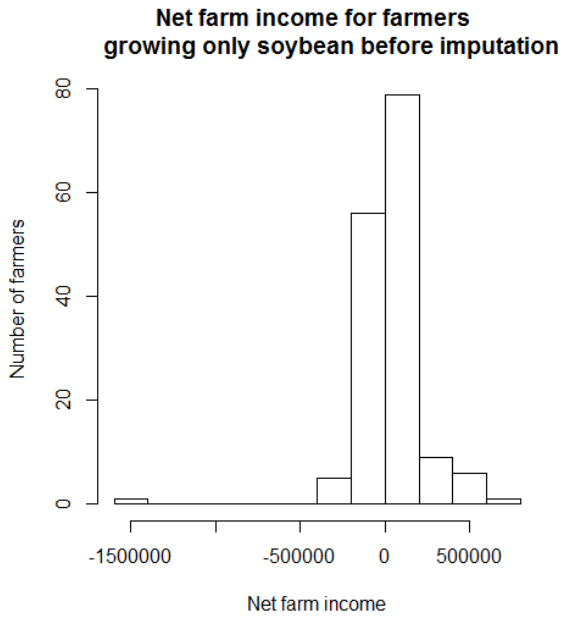


Figure 5.5 Wheat Total Net Return Imputation

(a) Soybean Total Net Return before Imputation



(a) Soybean Total Net Return after Imputation

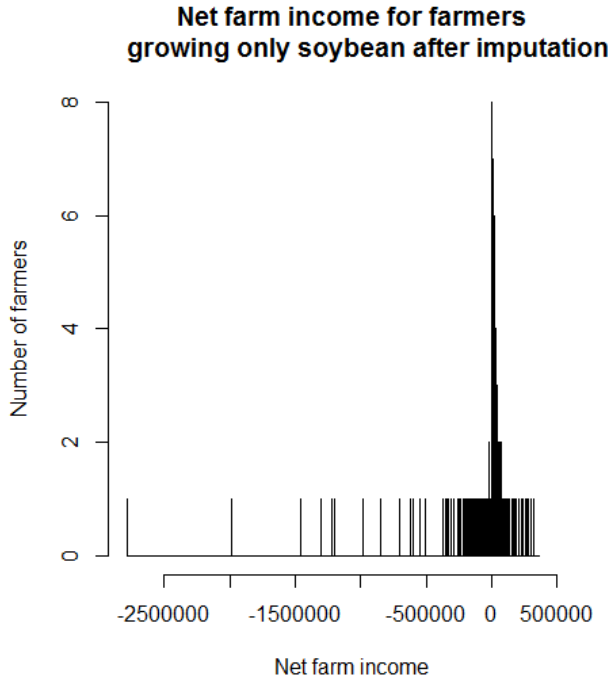
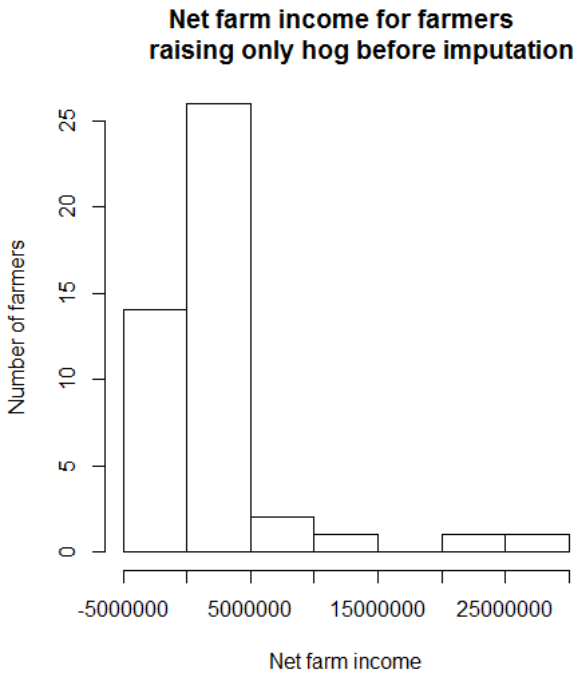


Figure 5.6 Soybean Total Net Return Imputation

(a) Hog Total Net Return before Imputation



(a) Hog Total Net Return after Imputation

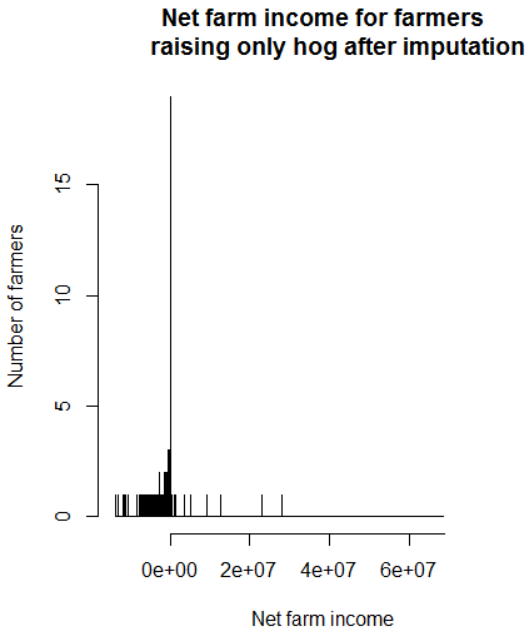
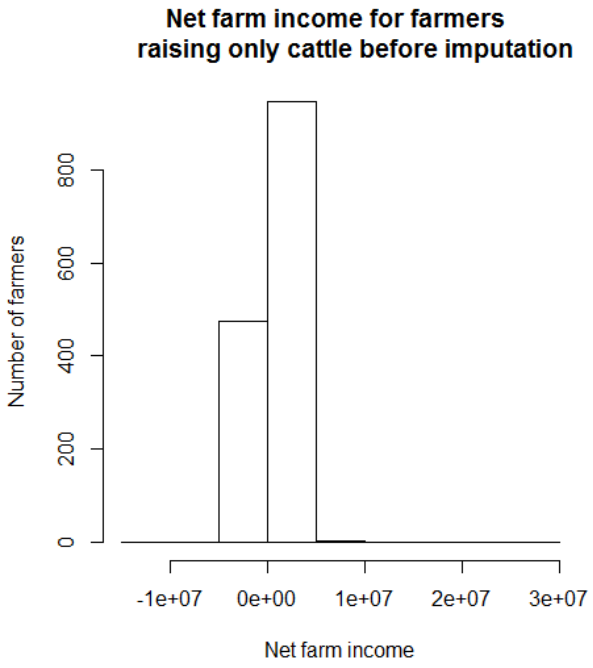


Figure 5.7 Hog Total Net Return Imputation

(a) Cattle Total Net Return before Imputation



(b) Cattle Total Net Return after Imputation

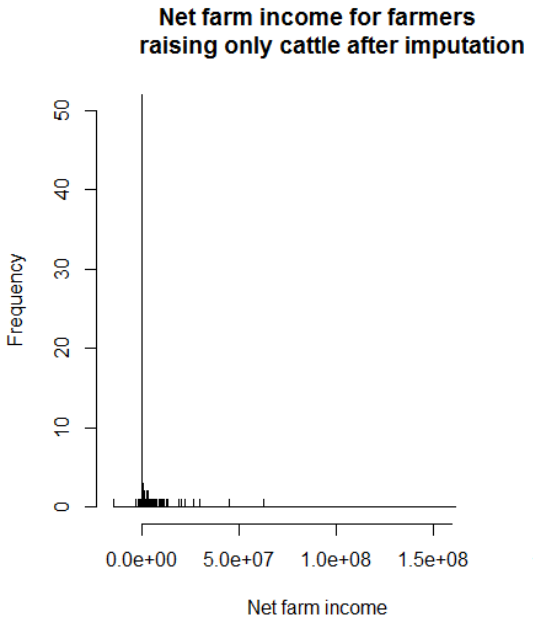
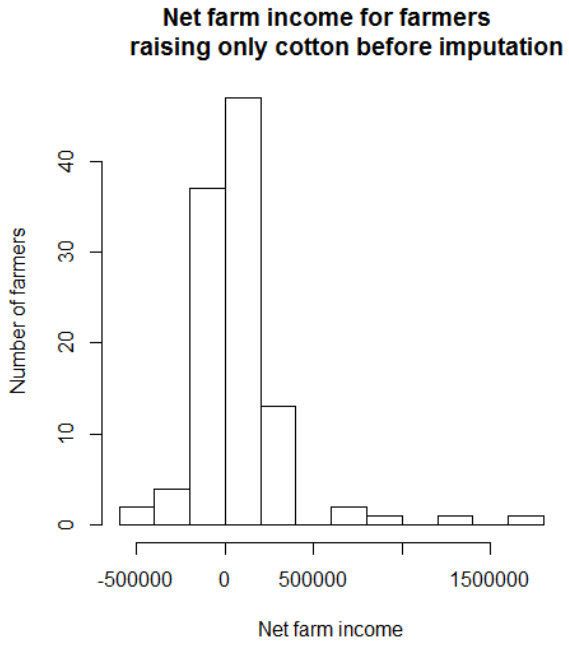


Figure 5.8 Cattle Total Net Return Imputation

(a) Cotton Net Return before Imputation



(b) Cotton Net Return after Imputation

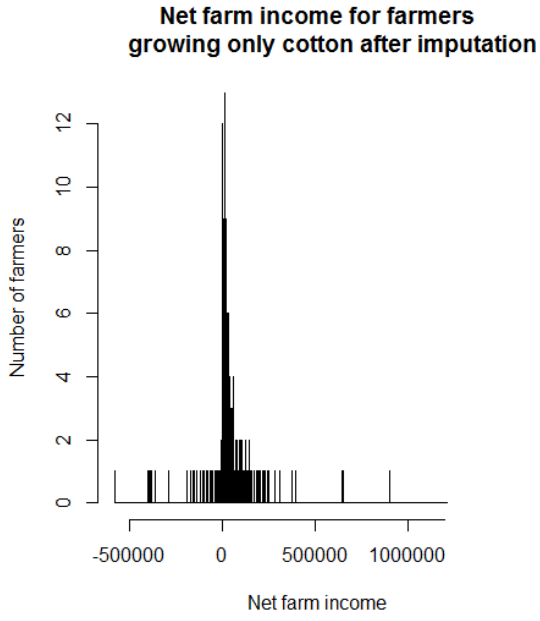


Figure 5.9 Cotton Total Net Return Imputation

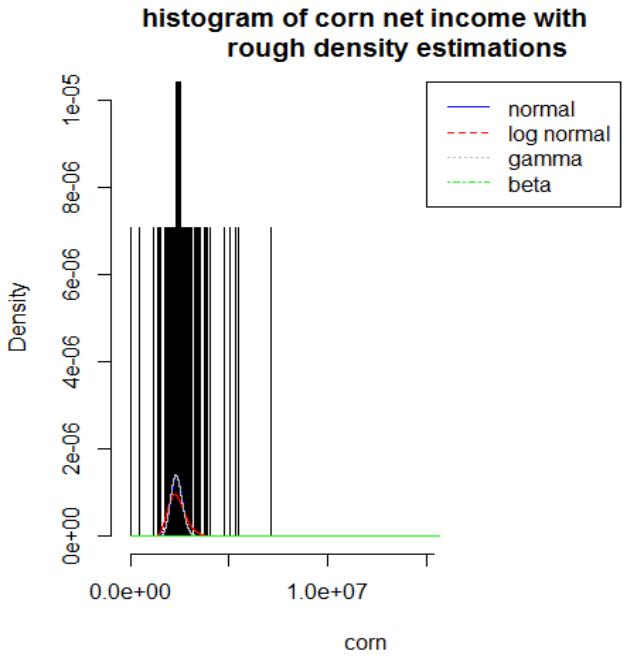


Figure 5.10 Fit Different Distributions to Imputed Corn Total Net Return

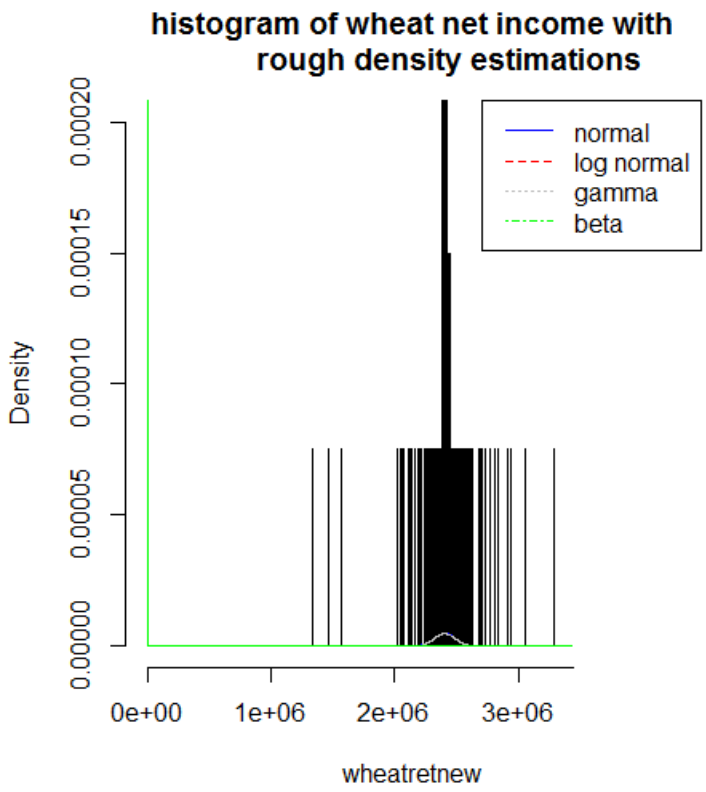


Figure 5.11 Fit Different Distributions to Imputed Wheat Total Net Return

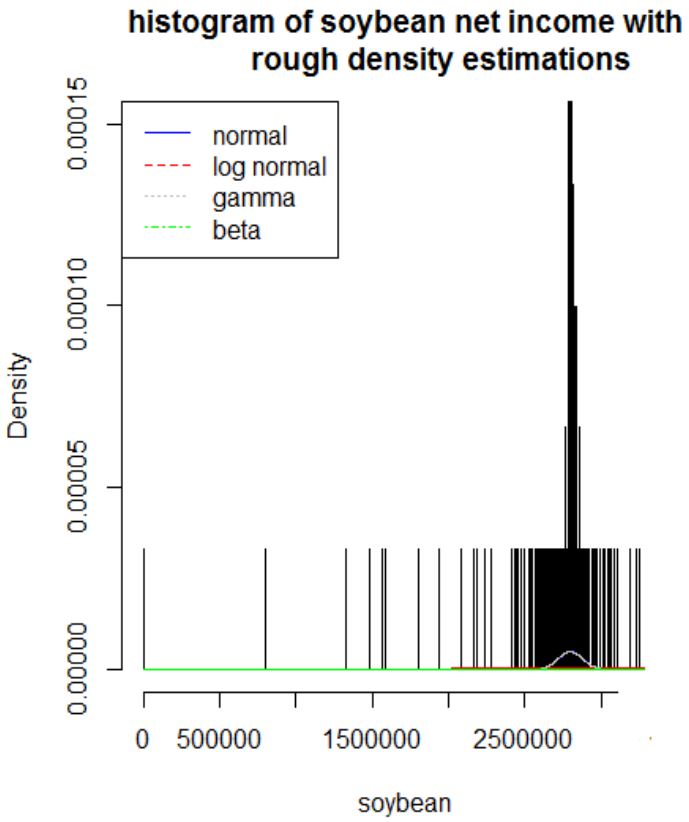


Figure 5.12 Fit Different Distributions to Imputed Soybean Total Net Return

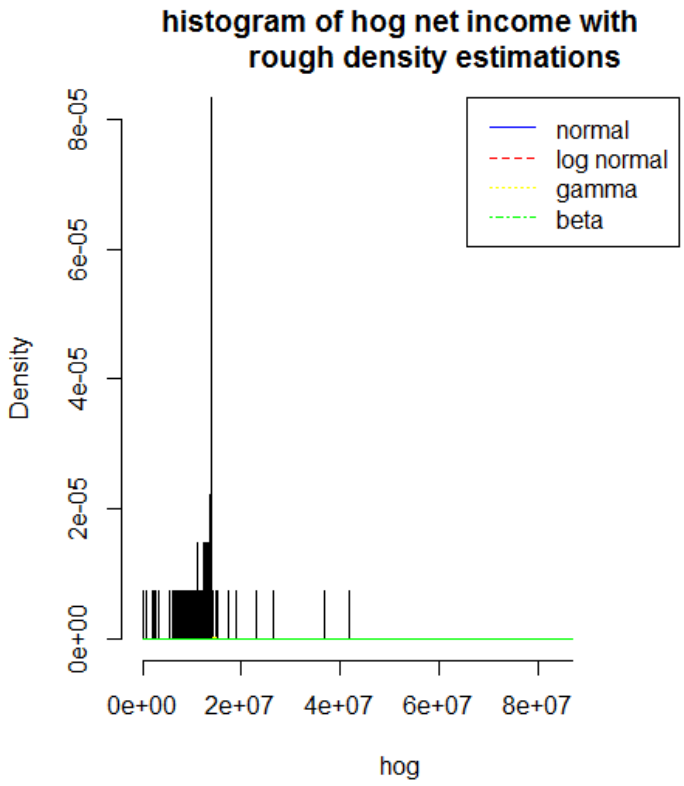


Figure 5.13 Fit Different Distributions to Imputed Hog Total Net Return

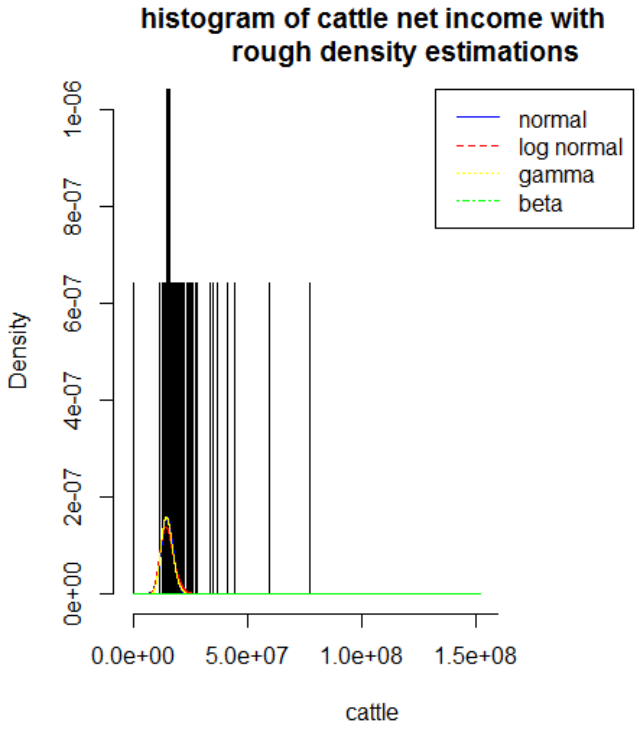


Figure 5.14 Fit Different Distributions to Imputed Cattle Total Net Return

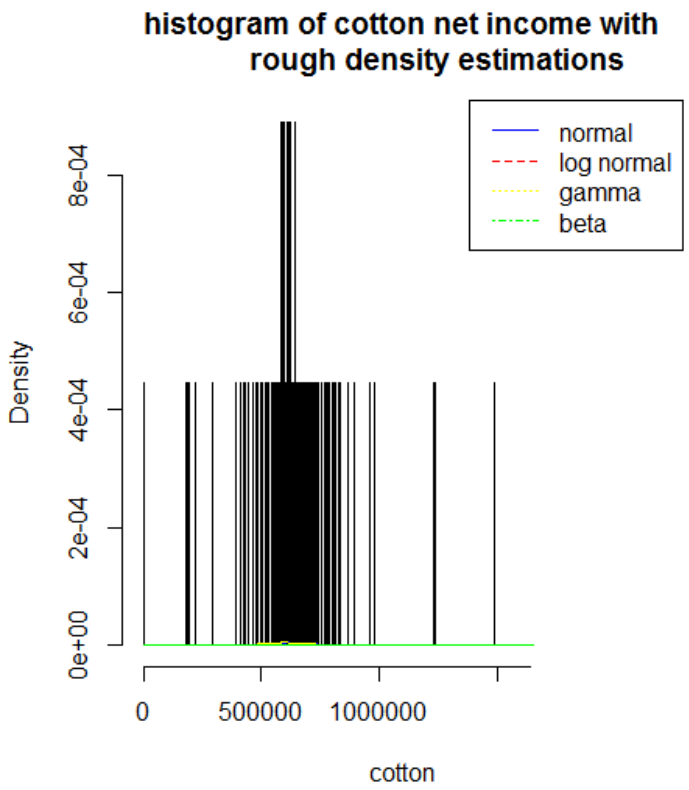


Figure 5.15 Fit Different Distributions to Imputed Cotton Total Net Return

Table 5.1 Different Proportions of T and F with Different Base Line at Different Missing Proportions

	proportion below 1e-04	**proportion below 1e-03**	**proportion below 1e-02**	**proportion below 1e-01**
missing proportion(F)0.205	0.474817073171	0.474780487805	0.474243902439	0.470695121951
missing proportion(T)0.205	0.525182926829	0.525219512195	0.525756097561	0.529304878049
missing proportion(F)0.225	0.474011111111	0.473988888889	0.473544444444	0.469577777778
missing proportion(T)0.225	0.525988888889	0.526011111111	0.526455555556	0.530422222222
missing proportion(F)0.25	0.477050000000	0.477000000000	0.476750000000	0.472940000000
missing proportion(T)0.25	0.522950000000	0.523000000000	0.523250000000	0.527060000000
missing proportion(F)0.275	0.475345454545	0.475318181818	0.475027272727	0.471236363636
missing proportion(T)0.275	0.524654545455	0.524681818182	0.524972727273	0.528763636364
missing proportion(F)0.3	0.476625000000	0.476583333333	0.476133333333	0.472366666667
missing proportion(T)0.3	0.523375000000	0.523416666667	0.523866666667	0.527633333333
missing proportion(F)0.325	0.477261538462	0.477223076923	0.476907692308	0.473292307692
missing proportion(T)0.325	0.522738461538	0.522776923077	0.523092307692	0.526707692308
missing proportion(F)0.35	0.474464285714	0.474435714286	0.474207142857	0.470614285714
missing proportion(T)0.35	0.525535714286	0.525564285714	0.525792857143	0.529385714286
missing proportion(F)0.375	0.477333333333	0.477293333333	0.476866666667	0.473253333333
missing proportion(T)0.375	0.522666666667	0.522706666667	0.523133333333	0.526746666667
missing proportion(F)0.4	0.479562500000	0.479518750000	0.479225000000	0.475850000000
missing proportion(T)0.4	0.520437500000	0.520481250000	0.520775000000	0.524150000000
missing proportion(F)0.425	0.477835294118	0.477811764706	0.477529411765	0.474252941176
missing proportion(T)0.425	0.522164705882	0.522188235294	0.522470588235	0.525747058824
missing proportion(F)0.45	0.477644444444	0.477622222222	0.477266666667	0.473733333333
missing proportion(T)0.45	0.522355555556	0.522377777778	0.522733333333	0.526266666667
missing proportion(F)0.475	0.476321052632	0.476268421053	0.475931578947	0.472584210526
missing proportion(T)0.475	0.523678947368	0.523731578947	0.524068421053	0.527415789474
missing proportion(F)0.5	0.479695000000	0.479690000000	0.479445000000	0.476545000000
missing proportion(T)0.5	0.520305000000	0.520310000000	0.520555000000	0.523455000000
missing proportion(F)0.525	0.479200000000	0.479171428571	0.478895238095	0.475547619048
missing proportion(T)0.525	0.520800000000	0.520828571429	0.521104761905	0.524452380952
missing proportion(F)0.55	0.478754545455	0.478731818182	0.478481818182	0.475286363636
missing proportion(T)0.55	0.521245454545	0.521268181818	0.521518181818	0.524713636364
missing proportion(F)0.575	0.478030434783	0.478004347826	0.477691304348	0.474656521739
missing proportion(T)0.575	0.521969565217	0.521995652174	0.522308695652	0.525343478261
missing proportion(F)0.6	0.477458333333	0.477395833333	0.477075000000	0.474079166667
missing proportion(T)0.6	0.522541666667	0.522604166667	0.522925000000	0.525920833333
missing proportion(F)0.625	0.480448000000	0.480416000000	0.480120000000	0.477084000000
missing proportion(T)0.625	0.519552000000	0.519584000000	0.519880000000	0.522916000000
missing proportion(F)0.65	0.478757692308	0.478726923077	0.478415384615	0.475353846154
missing proportion(T)0.65	0.521242307692	0.521273076923	0.521584615385	0.524646153846
missing proportion(F)0.675	0.482466666667	0.482448148148	0.482140740741	0.478903703704
missing proportion(T)0.675	0.517533333333	0.517551851852	0.517859259259	0.521096296296
missing proportion(F)0.7	0.481060714286	0.480992857143	0.480678571429	0.477932142857
missing proportion(T)0.7	0.518939285714	0.519007142857	0.519321428571	0.522067857143

Table 5.2 Summary Statistics of Per Acre Net Return for Different Commodities in Simulated Farm Data

	mean	Standard Deviation	Median
corn	26.72196	70.2881711	0
wheat	6.743689	53.9085429	0
soybean	42.36286	69.6558684	12.54183

Table 5.3 Mean Absolute Percentage Error on Commodity Net Return in Simulated Farm Data at Different Missing Proportion after Missing Data Recovery

Missing Proportion	MAPE (%)	Missing Proportion	MAPE (%)	Missing Proportion	MAPE (%)
0.2	3.923766429	0.465	2.372655505	0.73	0.129536642
0.205	0.581173480	0.47	1.227080468	0.735	6.866357866
0.21	3.780938567	0.475	8.922851439	0.74	8.149283570
0.215	6.876860466	0.48	13.364547859	0.745	0.485150853
0.22	2.420462421	0.485	1.636019445	0.75	16.023853463
0.225	0.573391756	0.49	11.420353059	0.755	21.040615877
0.23	3.033496699	0.495	8.877220755	0.76	7.831495449
0.235	3.888182639	0.5	8.618287421	0.765	12.126011362
0.24	9.580822340	0.505	21.077539272	0.77	8.962670199
0.245	9.602580005	0.51	3.122146405	0.775	1.236906849
0.25	2.917457991	0.515	10.128003064	0.78	2.560110381
0.255	2.116221294	0.52	6.317763595	0.785	11.887299553
0.26	3.438768369	0.525	10.059460858	0.79	2.194682022
0.265	1.691481605	0.53	10.434588499	0.795	14.663092610
0.27	2.342269877	0.535	6.051707629	0.8	2.078572561
0.275	5.078849770	0.54	5.562434882	0.805	1.988848054
0.28	3.784718920	0.545	11.356006915	0.81	1.889148667
0.285	5.107335502	0.55	5.495236026	0.815	3.108192421
0.29	5.315727849	0.555	0.068052538	0.82	10.561532843
0.295	2.400821472	0.56	16.074637553	0.825	2.082151189
0.3	11.890081350	0.565	10.022765792	0.83	10.342473823
0.305	10.271542461	0.57	3.047374631	0.835	0.025541650
0.31	10.898251306	0.575	9.785777483	0.84	5.733658056
0.315	0.266487805	0.58	14.314255739	0.845	0.929737279
0.32	9.413308885	0.585	0.223516612	0.85	17.292792289
0.325	2.437380527	0.59	17.830428031	0.855	4.352087396
0.33	1.712777474	0.595	16.132891509	0.86	6.598723628
0.335	13.698915330	0.6	0.601746345	0.865	14.223016660
0.34	0.988357651	0.605	1.728207368	0.87	12.810257885
0.345	1.468545574	0.61	5.534661121	0.875	12.715585945
0.35	1.326759611	0.615	4.040532322	0.88	23.983422202
0.355	2.346993326	0.62	4.180725047	0.885	28.168751791
0.36	7.701648194	0.625	12.183948833	0.89	13.815456214
0.365	1.547505945	0.63	2.095166775	0.895	10.052776476
0.37	0.344962657	0.635	2.453521871	0.90	12. 33986342
0.375	20.590469096	0.64	17.487751947		
0.38	0.352731595	0.645	9.648481731		
0.385	4.142904536	0.65	5.799517448		
0.39	9.192838411	0.655	4.837959869		
0.395	0.517543718	0.66	2.343104139		
0.4	5.447399492	0.665	14.551246502		
0.405	12.074014479	0.67	3.619048354		
0.41	12.825612238	0.675	21.017406916		
0.415	11.357940317	0.68	10.832590144		
0.42	5.917049243	0.685	5.992972179		
0.425	6.608499876	0.69	9.464292803		
0.43	6.964289505	0.695	13.887689011		
0.435	3.501222424	0.7	12.819733217		
0.44	0.496334506	0.705	18.852776645		
0.445	4.898930317	0.71	8.211810756		
0.45	0.073192660	0.715	15.985168188		
0.455	3.300733265	0.72	16.037369151		
0.46	8.372525905	0.725	6.988012749		

Table 5.4 Mean Absolute Error on Commodity Net Return in Simulated Farm Data at Different Missing Proportion after Missing Data Recovery

Missing Proportion	MAE	Missing Proportion	MAE	Missing Proportion	MAE
0.2	9.946773	0.465	21.784476	0.73	40.041343
0.205	9.924946	0.47	22.312451	0.735	36.776945
0.21	10.237093	0.475	21.751151	0.74	40.235623
0.215	10.205330	0.48	24.852341	0.745	42.113848
0.22	10.455178	0.485	22.696616	0.75	50.782610
0.225	10.461050	0.49	21.771552	0.755	40.880853
0.23	10.696942	0.495	23.762174	0.76	42.633478
0.235	11.365137	0.5	26.397458	0.765	50.055588
0.24	11.811590	0.505	25.449714	0.77	43.023367
0.245	11.131155	0.51	26.267179	0.775	46.260535
0.25	11.978321	0.515	25.089621	0.78	42.305448
0.255	11.965178	0.52	27.201251	0.785	38.771664
0.26	12.512153	0.525	29.877147	0.79	42.197636
0.265	12.400833	0.53	27.568592	0.795	62.288909
0.27	12.784236	0.535	25.755020	0.8	42.538546
0.275	13.484570	0.54	26.012057	0.805	48.282243
0.28	12.833401	0.545	27.974059	0.81	47.214184
0.285	13.638507	0.55	51.344249	0.815	44.086463
0.29	14.534820	0.555	27.681542	0.82	50.288359
0.295	14.599279	0.56	28.712476	0.825	52.032001
0.3	14.392001	0.565	30.444877	0.83	47.345707
0.305	14.860034	0.57	29.176483	0.835	56.124562
0.31	15.085416	0.575	29.511068	0.84	52.869036
0.315	15.542744	0.58	37.979245	0.845	48.829715
0.32	15.371723	0.585	37.469913	0.85	48.923798
0.325	15.604334	0.59	35.746951	0.855	53.226322
0.33	15.794171	0.595	31.196056	0.86	50.314066
0.335	16.427177	0.6	30.150422	0.865	52.805876
0.34	16.185912	0.605	30.674223	0.87	49.534088
0.345	17.096472	0.61	31.722152	0.875	57.874651
0.35	16.444755	0.615	31.222069	0.88	53.445176
0.355	17.097407	0.62	32.225838	0.885	54.099719
0.36	16.512288	0.625	32.838512	0.89	55.687268
0.365	16.911772	0.63	34.088430	0.895	56.117335
0.37	17.670896	0.635	32.720153	0.90	56.072358
0.375	18.409156	0.64	35.743376		
0.38	18.995859	0.645	30.359980		
0.385	18.889473	0.65	31.456661		
0.39	17.738961	0.655	35.590602		
0.395	18.433338	0.66	32.319150		
0.4	18.457787	0.665	38.020011		
0.405	20.035867	0.67	33.379680		
0.41	20.082503	0.675	41.329075		
0.415	20.001476	0.68	32.587661		
0.42	21.163655	0.685	34.848134		
0.425	28.397695	0.69	35.466615		
0.43	20.099547	0.695	33.125630		
0.435	19.935461	0.7	37.677277		
0.44	21.478250	0.705	35.072963		
0.445	22.742858	0.71	40.258434		
0.45	22.080054	0.715	40.279253		
0.455	21.352318	0.72	47.832634		
0.46	20.968295	0.725	40.862582		

Table 5.5 Test Statistics on Fitting Potential Distributions to Commodity-Specific Per Acre Net Return in Simulated Farm Data

	Normal Distribution (p-value)	Weibull Distribution (p-value)	Gamma Distribution (p-value)	Beta Distribution (p-value)
Corn	0.3461 (<2.2e-16)	NA	0.0848 (<2.2e-16)	0.0791 (<2.2e-16)
Soybean	0.0614 (<2.2e-16)	NA	0.1163 (<2.2e-16)	0.1216 (<2.2e-16)
Wheat	0.3202 (<2.2e-16)	NA	0.3461 (<2.2e-16)	0.3638 (<2.2e-16)

Table 5.6 Basic Statistics on Total Net Return of Main Commodities in ARMS after Missing Data Recovery

	Maximum Total Net Return	Average Total Net Return for Each Farmer
Corn	25161020	55189.84
Wheat	1918708	20211.84
Soybean	658839	12607.87
Hog	40698818	-939629.2
Cattle	19573855	190320.1
Cotton	1614009	27330.96

Table 5.7 Two Sample K-S Test on Net Return of Main Commodities before and after Missing Data Recovery

	K-S Statistics	p-value
Corn	0.89179	<2.2e-16
Wheat	0.46719	<2.2e-16
Soybean	0.45082	<2.2e-16
Hog	0.90361	<2.2e-16
Cattle	0.36878	<2.2e-16
Cotton	0.53704	<2.2e-16

Table 5.8 Basic Statistics on Middle Part of Per Acre Net Return or Per Animal Net Return after Missing Data Recovery

	mean	variance
corn	73.95251	830.3979
wheat	45.27663	276.1955
soybean	50.74948	135.6001
hog	4.521242	8.536004
cattle	117.1832	742.6035
cotton	53.22941	802.6128

Table 5.9 Quantile Estimation on Middle Part of Per Acre Net Return or Per Animal Net Return after Missing Data Recovery

	5% in distribution	15% in distribution	25% in distribution	50% in distribution	75% in distribution	90% in distribution	95% in distribution
corn	34.3494	62.6055	68.2669	71.5273	75.9362	78.9579	82.3450
wheat	40.1725	43.1322	44.6478	46.3830	47.4783	49.4207	51.0167
soybean	45.2162	49.8312	50.2974	51.4323	52.0803	54.8903	56.2227
Hog	0.0485	0.5447	1.9136	5.8169	6.3674	7.5117	7.82230
cattle	77.2391	109.9143	110.2257	113.1276	121.7811	139.5000	158.0912
cotton	32.4071	42.8562	47.8322	53.8982	55.3929	56.8152	60.0605

Table 5.10 Kullback-Leibler Divergence for Different Kernel Density Estimation on Commodity Net Return

	Normal Distribution	Log Normal Distribution	Weibull Distribution	Gamma Distribution	Beta Distribution
Corn	0.7562488	Inf	Inf	Inf	inf
Wheat	35.80676	0.4378197	inf	inf	inf
Soybean	52.73173	inf	NA	inf	inf
Hog	0.04500874	0.2921279	NA	0.7142646	Inf
Cattle	0.04545684	Inf	NA	0.3098259	inf
Cotton	5.942488	27.947	NA	55.24141	inf

References

Andridge, R. R., and Roderick J.A. Little (2010). A Review of Hot Deck Imputation for Survey Non-Response. International Statistical Review (2010) Volume 78 Issue 1, 40-60.

Athey, S., Bayati, M., Doudchenko, N. and Imbens, G. (2017). Matrix Completion Methods for Causal Panel Data Models. Working Paper (2017), retrieved form https://arxiv.org/abs/1710.10251.

Broeck, G. V., Mohan, K., Choi, A. and Pearl, J. (2014). Efficient Algorithms for Bayesian Network Parameter Learning from Incomplete Data. Thirty-First Conference on Uncertainty in Artificial Intelligence (2014), 161-170.

Basturk, N., and Cakmakli, C., Ceyhan, C. P. and Dijk, H. K. (2013). Historical Development in Bayesian Econometrics after Cowles Foundation Monographs. Tinbergen Institute Discussion Paper (2013).

Candès, E. J., and Plan, Y. (2010). Matrix Completion with Noise. Proceedings of IEEE (2010) Volume 98 Issue 6, 925-936.

Candès, E. J., and Recht, B. (2009). Exact Matrix Completion via Convex Optimizaiton. Foundations of Computational Mathematics (2009) Volume 9 Issue 6, 111-119.

Candès, E. J., Romberg, J. and Tao, T. (2006). Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. IEEE Transaction on Information Theory (2006) Volume 52 Issue 2, 489-509.

Candès, E. J., and Tao, T. (2010). The Power of Convex Relaxation: Near-Optimal Matrix Completion. IEEE Transaction on Information Theory (2010) Volume 56 Issue 5, 2053-2080.

Cai, J. F., Candès, E. J. and Zuowei Shen (2010). A singular Value Thresholding Algorithm for Matrix Completion. SIAM Journal on Optimizaiton (2010) Volume 20 Issue 4, 1956-1982.

Cai, T., T. Tony Cai and Zhang, A. (2016). Structured Matrix Completion with Applications to Genomic Data Integration. Journal of the American Statistical Association (2016) Volume 111 Issue 514, 621-633.

Hastie, T. J., Mazumder, T., Lee, J. D. and Zadeh, R. B. (2015). Matrix completion and low-rank SVD via fast alternating least Squares. Journal of Machine Learning Research (2015) 3367-3402.

Linero, A. R. and Danielsy, M. J. (2018). Bayesian Approaches for Missing Not at Random Outcome Data: The Role of Identifying Restrictions. Statistical Science (2018), Volume 33 Number 2, 198-213.

Little, R. (2011). Calibrated Bayes, for Statistics in General, and Missing Data in Particular. Statistical Science (2011), Volume 26 Issue 2, 162-174.

Keshavan, R. H., Montanari, A. and Sewoong Oh (2010). Matrix Completion from Noise Entries. Journal of Machine Learning Research (2010), 2057-2078.

Keshavan, R. H. and Andrea Montanari (2010). Regularization for Matrix Completion. 2010 IEEE International Symposium on Information Theory. DOI: 10.1109/ISIT.2010.5513563

Krishnamurthy, Akshay (2011). Some Properties of Matrix Norms. Class notes in CMU (2011).

Mitra, Robin (2008). Bayesian Methods to Impute Missing Covariates for Causal Inference and Model Selection (Doctoral dissertation). https://pdfs.semanticscholar.org/dd83/1225a76a7fbc61ab3f64ab364532eea2b750.pdf.

MIYAKOSHI, Y. and KATO, S. (2012). A Missing Value Imputation Method Using a Bayesian Network with Weighted Learning. Electronics and Communications in Japan (2012), Volume 95 Issue 12, 299-305.

Peng, X., Lu, C., Yi, Z. and Tang, H. (2016). Connections Between Nuclear Norm and Frobenius Norm Based Representations. IEEE Trans. on Neural Networks and Learning Systems (2016) 218-224.

Recht, B., (2011). A Simpler Approach to Matrix Completion. Journal of Machine Learning Research (2011), 3413-3430.

Recht, B., M. Fazel and P. A. Parrilo (2010). Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minizaiton. SIAM Review (2010) Volume 52 Issue 3, 471-571.

Schmitt, P., J. Mandel and M. Guedj (2015). A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics & Biostatistics (2015) Volume 6 Issue 1, 224-230.

Smith, M.J. De (2014). Statistical Analysis Handbook (2014), 12-5.

Srebo, N., Jason, R. and Jaakkola, T. (2005). Maximum Margin Matrix Factorizaiton. Advances in Neural Information Progressing Systems (2005) Volume 17, 1-8.

The Agricultural Resource Management Survey (ARMS) database is USDA's primary source of information on the financial condition, production practices, and resource use of America's farm businesses and the farm households. Retrieved July 24, 2017, from https://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices/

Zhou, X. and Reiter, J. P. (2010). A Note on Bayesian Inference After Multiple Imputation. The American Statistician (2010), Volume 64 Issue 2, 159-163.

Zio, M. D., Scanu, M., Coppola, L., Luzi, O. and Ponti, A. (2004). Bayesian networks for imputation. Journal of the Royal Statistical Society (2004), Volume 167 Issue 2, 309-322.

Appendices

Appendix A

Proof for some theorems and lemmas (All detailed proof work for all lemmas and theorems are done by Candès, E. J., and B. Recht (2009) and Cai et al. (2008))

Proof for lemma 1

The bound on $p^{-1}\|P_{T^\perp}P_\Omega P_{\mathrm{T}}(\boldsymbol{W})\| = p^{-1}\|P_{T^\perp}(P_\Omega - p\mathrm{I})P_{\mathrm{T}}(\boldsymbol{W})\|$ is equal to $p^{-1}\|(P_\Omega - p\mathrm{I})\boldsymbol{W}\|$. We should notice the facts that $P_{T^\perp}P_{\mathrm{T}} = 0$ and $\|P_{T^\perp}(\boldsymbol{W})\| \leq \|\boldsymbol{W}\|$ are valid for any matrix $\boldsymbol{W}$ in our case. Define $\boldsymbol{R} \equiv p^{-1}(P_\Omega - p\mathrm{I})(\boldsymbol{W})$=$p^{-1}\sum_{ab}(\delta_{ab} - p)\, E_{ab}e_a e_b^*$. $\boldsymbol{R}$ depends on δ_{ab}. Both of δ_{ab} and $\boldsymbol{R}$ are random variables. The function $f(\boldsymbol{R}) = \|\boldsymbol{R}\|^q$ is convex, Jensen's inequality leads to that $\mathbb{E}\|\boldsymbol{R}\|^q \leq \mathbb{E}\|\boldsymbol{R} - \boldsymbol{R}'\|^q$, where $\mathbb{E}\,\boldsymbol{R} = 0$ and $\boldsymbol{R}' = p^{-1}\sum_{ab}(\delta'_{ab} - p)\, E_{ab}e_a e_b^*$. Because $(\delta_{ab} - \delta'_{ab})$ is symmetric and $\boldsymbol{R}$ shares the same distribution as $\boldsymbol{R}'$, we can get that $p^{-1}\sum_{ab}\epsilon_{ab}(\delta_{ab} - \delta'_{ab})\, E_{ab}e_a e_b^* \equiv \boldsymbol{R}_\epsilon - \boldsymbol{R}'_\epsilon$, where $\{\epsilon_{ab}\}$ is Rademacher sequence and $\boldsymbol{R}_\epsilon$ = $p^{-1}\sum_{ab}\epsilon_{ab}\delta_{ab}\, E_{ab}e_a e_b^*$. Based on the triangle inequality, we derive:

$(\mathbb{E}\|\boldsymbol{R}_\epsilon - \boldsymbol{R}'_\epsilon\|^q)^{1/q} \leq (\mathbb{E}\|\boldsymbol{R}_\epsilon\|^q)^{1/q} + (\mathbb{E}\|\boldsymbol{R}'_\epsilon\|^q)^{1/q} = 2(\mathbb{E}\|\boldsymbol{R}_\epsilon\|^q)^{1/q}$ and

$(\mathbb{E}\|\boldsymbol{R}\|^q)^{1/q} \leq 2p^{-1}(\mathbb{E}_\delta\mathbb{E}_\epsilon\|\sum_{ab}\epsilon_{ab}\delta_{ab}\, E_{ab}e_a e_b^*\|^q)^{1/q}$.

For any q greater than 1, the Schatten q-norm of one matrix is defined as:

$\|\boldsymbol{W}\|_{s_q} = (\sum_{i=1}^{n}\sigma_i(\boldsymbol{W})^q)^{1/q}$. The Frobenius-norm is the Schatten 2-norm and the nuclear norm is Schatten 1-norm. Combining with work from Buchholz and Lust-Picquard, our proof work is done.

Proof for Theorem 1 (Candès, E. J., and B. Recht, 2009**)**

Any matrix $\boldsymbol{W}$ can be decomposed as $\boldsymbol{W} = \sum_{ab}\langle \boldsymbol{W}, e_a e_b^*\rangle\, e_a e_b^*$, this gives

$P_{\mathrm{T}}(\boldsymbol{W}) = \sum_{ab}\langle P_{\mathrm{T}}(\boldsymbol{W}), e_a e_b^*\rangle\, e_a e_b^* = \sum_{ab}\langle \boldsymbol{W}, P_{\mathrm{T}}(e_a e_b^*)\rangle\, e_a e_b^*$. Therefore, $P_\Omega P_{\mathrm{T}}(\boldsymbol{W}) = \sum_{ab}\gamma_{ab}\langle \boldsymbol{W}, P_{\mathrm{T}}(e_a e_b^*)\rangle\, e_a e_b^*$ which can be derived as:

$(P_{\mathrm{T}}P_\Omega P_{\mathrm{T}})(\boldsymbol{W}) = \sum_{ab}\gamma_{ab}\langle \boldsymbol{W}, P_{\mathrm{T}}(e_a e_b^*)\rangle\, P_{\mathrm{T}}(e_a e_b^*)$. In other words, $P_{\mathrm{T}}P_\Omega P_{\mathrm{T}} = \sum_{ab}\gamma_{ab}\, P_{\mathrm{T}}(e_a e_b^*) \otimes P_{\mathrm{T}}(e_a e_b^*)$

Proof for Lemma 2 (Candès, E. J., and B. Recht, 2009**)**

By bounding the spectral norm of $P_{T^\perp}P_\Omega P_{\mathrm{T}}H(\boldsymbol{W})$, we have the following inequality based on the previous knowledge:

$p^{-1}\|(P_{T^\perp}P_\Omega P_{\mathrm{T}})\boldsymbol{\Lambda}(\boldsymbol{W})\| \le p^{-1}\|(P_\Omega - pI)\boldsymbol{\Lambda}(\boldsymbol{W})\|$,

If we define $\boldsymbol{N}$=$p^{-1}(P_\Omega - pI)\boldsymbol{\Lambda}(\boldsymbol{W}) = p^{-2}\sum_{ab,a'b'}\zeta_{ab}\zeta_{,a'b'}E_{a'b'}\,\langle P_T e_{a'}e_{b'}^*, e_a e_b^*\rangle e_a e_b^*$,

where $\zeta_{ab} \equiv \gamma_{ab} - p$. We decompose $\boldsymbol{N}$ into $\boldsymbol{N} \equiv \boldsymbol{N_0} + \boldsymbol{N_1}$.

Beginning with $\boldsymbol{N_0}$, we break $(\zeta_{ab})^2$ into $(1-2p)\zeta_{ab} + p(1-p)$. We can express $\boldsymbol{N_0}$ as

$$\boldsymbol{N_0} = \frac{1-2p}{p}\sum_{ab}\zeta_{ab}\boldsymbol{\Lambda}_{ab}\, e_a e_b^* + (1-p)\sum_{ab}\boldsymbol{\Lambda}_{ab}\, e_a e_b^*, H_{ab} \equiv p^{-1}W_{ab}\langle P_T e_a e_b^*, e_a e_b^*\rangle$$

The spectral norm of first term is bounded. We can get

$p^{-1}\|\zeta_{ab}H_{ab}e_a e_b^*\| \le C_0\sqrt{\frac{n^3(\beta\log n)}{m}}\|\boldsymbol{\Lambda}\|_\infty$ with probability at least $1 - n^{-\beta}$. We know that

$\|\boldsymbol{W}\|_\infty \le \mu_1\sqrt{r}/n$, $\|\boldsymbol{\Lambda}\|_\infty \le \mu_0\mu_1(2r/np)\sqrt{r}/n$ and

$p^{-1}\|\sum_{ab}\zeta_{ab}H_{ab}\, e_a e_b^*\| \le C\mu_0\mu_1\frac{nr}{m}\sqrt{\frac{nr(\beta\log n)}{m}}$ hold with the same probability.

Proof for Lemma 3 (Candès, E. J., and B. Recht, 2009**)**

Lemma 3 is very close to Lemma 2 Its proof is also similar to the proof of Lemma 3.

Proof for Lemma 4 (Candès, E. J., and B. Recht, 2009**)**

Here, we can continue our proof work in the same direction to estimate the spectral norm of $p^{-1}(P_\Omega - pI)\boldsymbol{\Lambda}^3(\boldsymbol{W})$ based on the previous work. In other words, $p^{-1}(P_\Omega - pI)\boldsymbol{\Lambda}^3(\boldsymbol{W})$ can be expressed as:

$p^{-4}\sum_{\omega_1,\omega_2,\omega_3,\omega_4}\left[\prod_{i=1}^{4}\zeta_{\omega_i}\right]E_{\omega_4}\left[\prod_{i=1}^{3}P_{\omega_i+1\omega_i}\right]\boldsymbol{F}_{\omega_1}$. The notations are the same as previous section. The sum in expression of $p^{-1}(P_\Omega - pI)H^3(\boldsymbol{W})$ depends on whether the $\omega_i{}'$s are the same or not. To bound the sum of every term, we use decoupling argument. This is easy to achieve with high possibility, but people need to consider 18 potential cases and calculate a lot. In this section, we propose the upper limit of the term $p^{-1}(P_\Omega - pI)\boldsymbol{\Lambda}^3(\boldsymbol{W})$. There are two advantages of such argument. The first one is that the argument is a very short one. Secondly, this argument builds on the theorem established in previous section.

Proof for Lemma 5 (Candès, E. J., and B. Recht, 2009**)**

When we study the spectral norm of $p^{-1}(P_{T^\perp}P_\Omega P_{\mathrm{T}})\sum_{k\geq k_0}\boldsymbol{\Lambda}^k(\boldsymbol{W})$ for some positive integer k_0, the bound of Frobenius norm is defined as:

$$p^{-1}\left\|(P_{T^\perp}P_\Omega P_{\mathrm{T}})\sum_{k\geq k_0}\boldsymbol{\Lambda}^k(\boldsymbol{W})\right\| \leq p^{-1}\left\|(P_\Omega P_{\mathrm{T}})\sum_{k\geq k_0}\boldsymbol{\Lambda}^k(\boldsymbol{W})\right\|_F$$

$$\leq \sqrt{3/2p}\left\|\sum_{k\geq k_0}\boldsymbol{\Lambda}^k(\boldsymbol{W})\right\|_F$$

where the inequality follows from previous conclusion. The upper limit of the Frobenius can be written as:

$$\left\|\sum_{k\geq k_0}\boldsymbol{\Lambda}^k(\boldsymbol{W})\right\|_F \leq \|\boldsymbol{\Lambda}\|^{k_0}\|\boldsymbol{W}\|_F + \|\boldsymbol{\Lambda}\|^{k_0+1}\|\boldsymbol{W}\|_F + \cdots$$

$$\leq \frac{\|\boldsymbol{\Lambda}\|^{k_0}}{1-\|\boldsymbol{\Lambda}\|}\|\boldsymbol{W}\|_F$$

Proof for lemma 7 (Cai, J. F., et al., 2009)

Any entry $\boldsymbol{U}$ of $\partial f_\varphi(\boldsymbol{Q})$ has the form that $\boldsymbol{U} = \varphi\boldsymbol{U}_0 + \boldsymbol{Q}$, where $\boldsymbol{U}_0 \in \partial\|\boldsymbol{Q}\|_*$. And $\boldsymbol{U}'$ holds the similar structure. We derive:

$$\langle \boldsymbol{U}-\boldsymbol{U}', \boldsymbol{Q}-\boldsymbol{Q}'\rangle = \varphi\langle \boldsymbol{U}_0-\boldsymbol{U}_0{}', \boldsymbol{Q}-\boldsymbol{Q}'\rangle + \|\boldsymbol{Q}-\boldsymbol{Q}'\|_F^2$$

The first term in right-hand side is nonnegative. Any subgradient of nuclear norm at $\boldsymbol{Q}$ follows $\|\boldsymbol{U}_0\|_2 \le 1$ and $\langle \boldsymbol{U}_0, \boldsymbol{Q}\rangle = \|\boldsymbol{Q}\|_*$. In addition, we find

$|\langle \boldsymbol{U}_0, \boldsymbol{Q}'\rangle| \le \|\boldsymbol{U}_0\|_2\|\boldsymbol{Q}'\|_* \le \|\boldsymbol{Q}'\|_*$, $|\langle \boldsymbol{U}_0, \boldsymbol{Q}\rangle| \le \|\boldsymbol{U}_0\|_2\|\boldsymbol{Q}\|_* \le \|\boldsymbol{Q}\|_*$

At the same time, it is reasonable to prove that $\langle \boldsymbol{U}_0-\boldsymbol{U}_0{}', \boldsymbol{Q}-\boldsymbol{Q}'\rangle$=$\langle \boldsymbol{U}_0, \boldsymbol{Q}\rangle$ + $\langle \boldsymbol{U}_0{}', \boldsymbol{Q}'\rangle$ − $\langle \boldsymbol{U}_0, \boldsymbol{Q}'\rangle - \langle \boldsymbol{U}_0{}', \boldsymbol{Q}\rangle$ = $\|\boldsymbol{Q}'\|_*$+$\|\boldsymbol{Q}\|_* - \langle \boldsymbol{U}_0, \boldsymbol{Q}'\rangle - \langle \boldsymbol{U}_0{}', \boldsymbol{Q}\rangle \ge 0$. Therefore, the lemma is proved.

Theorem 5 (Cai, J. F., et al., 2009)

If ($\boldsymbol{Q}^*$, $\boldsymbol{K}^*$) is primal-dual optimal pair for equation (12), by optimality condition, we can derive $\boldsymbol{U}$ satisfying that

$$\boldsymbol{U}^i - P_\Omega\left(\boldsymbol{K}^{i-1}\right) = \boldsymbol{0}$$
$$\boldsymbol{U}^* - P_\Omega(\boldsymbol{K}^*) = \boldsymbol{0},$$

for some $\boldsymbol{U}^k \in \partial f_\gamma(\boldsymbol{Q}^i)$ and some $\boldsymbol{U}^* \in \partial f_\gamma(\boldsymbol{Q}^*)$. Then, $\boldsymbol{U}$ deduces that

$$\left(\boldsymbol{U}^i - \boldsymbol{U}^*\right) - P_\Omega\left(\boldsymbol{K}^{i-1} - \boldsymbol{K}^*\right) = 0$$

and equations obeys that

$$\langle \boldsymbol{Q}^k - \boldsymbol{Q}^*, P_\Omega\left(\boldsymbol{K}^{i-1} - \boldsymbol{K}^*\right)\rangle = \langle \boldsymbol{U}^i - \boldsymbol{U}^*, \boldsymbol{Q}^i - \boldsymbol{Q}^*\rangle \ge \left\|\boldsymbol{Q}^i - \boldsymbol{Q}^*\right\|_F^2$$

Based on the observation, there should exist $P_\Omega\boldsymbol{Q}^* = P_\Omega\boldsymbol{W}$,

$\|P_\Omega(\boldsymbol{K}^k - \boldsymbol{K}^*)\|_F = \left\|P_\Omega\left(\boldsymbol{K}^{i-1} - \boldsymbol{K}^*\right) + \varphi_i P_\Omega\left(\boldsymbol{Q}^* - \boldsymbol{Q}^i\right)\right\|_F$.

If we assume $r_i = \left\|P_\Omega\left(\boldsymbol{K}^i - \boldsymbol{K}^*\right)\right\|_F$, we can get

$$r_i^2 = r_{i-1}^2 - 2\varphi_i\langle \boldsymbol{Q}^k - \boldsymbol{Q}^*, P_\Omega\left(\boldsymbol{K}^{i-1} - \boldsymbol{K}^*\right)\rangle + \varphi_i^2\left\|P_\Omega\left(\boldsymbol{Q}^* - \boldsymbol{Q}^i\right)\right\|_F^2$$
$$\le r_{i-1}^2 - 2\varphi_i\left\|\boldsymbol{Q}^i - \boldsymbol{Q}^*\right\|_F^2 + \varphi_i^2\left\|\boldsymbol{Q}^i - \boldsymbol{Q}^*\right\|_F^2$$

For any matrix $\boldsymbol{Q}$, $\|P_\Omega(\boldsymbol{Q})\|_F \le \|\boldsymbol{Q}\|_F$. Under our assumptions about the size of γ_k, we have $\varphi_i - \gamma_k^2 \ge \beta$ for all $k \ge 1$ and some $\beta > 0$ and thus

$r_k^2 \leq r_{k-1}^2 - \varphi_i^2 \left\| \boldsymbol{Q}^i - \boldsymbol{Q}^* \right\|_F^2$.

Proof for lemma 8 (Cai, J. F., et al., 2009)

Properties held by the projection $\boldsymbol{q}_0$ of one point $\boldsymbol{q}$ onto a convex set $\boldsymbol{C}$ is

$$\begin{cases} \boldsymbol{q}_0 \in \boldsymbol{C} \\ \langle \boldsymbol{k} - \boldsymbol{q}_0, \boldsymbol{q} - \boldsymbol{q}_0 \rangle \leq 0, \forall \boldsymbol{k} \in \boldsymbol{C} \end{cases}$$

where $\boldsymbol{C} = \mathbb{R}_+^m = \{\boldsymbol{q} \in \mathbb{R}^m : \boldsymbol{q} \geq \boldsymbol{0}\}$. Because $\boldsymbol{q}_0 \geq 0$,

$\langle \boldsymbol{k} - \boldsymbol{q}_0, \boldsymbol{q} - \boldsymbol{q}_0 \rangle \leq 0, \forall \boldsymbol{k} \geq 0$.

$\boldsymbol{k}^*$ is dual optimal, then we derive

$L(\boldsymbol{Q}^*, \boldsymbol{k}^*) \geq L(\boldsymbol{Q}^*, \boldsymbol{k}), \forall \boldsymbol{k} \geq 0$.

By substituting the expression in Lagrangian form,

$\langle \boldsymbol{k} - \boldsymbol{k}^*, F(\boldsymbol{Q}^*) \rangle \leq 0, \forall \boldsymbol{k} \geq 0$,

which is equivalent to

$\langle \boldsymbol{k} - \boldsymbol{k}^*, \boldsymbol{k}^* + \rho F(\boldsymbol{Q}^*) - \boldsymbol{k}^* \rangle \leq 0, \forall \boldsymbol{k} \geq 0, \forall \rho \geq 0$.

Thus, it is reasonable to conclude that $\boldsymbol{k}^*$ must be the projection of $\boldsymbol{k}^* + \rho F(\boldsymbol{Q}^*)$ onto the nonnegative orthant $\mathbb{R}_+^m$.

Proof for Lemma 9 (Cai, J. F., et al., 2009)

If $(\boldsymbol{Y}^*, \boldsymbol{k}^*)$ are primal-dual optimal pair for the equation (16). Based on the optimality condition, we know for all $\boldsymbol{Q}$:

$\langle \boldsymbol{U}^i, \boldsymbol{Q} - \boldsymbol{Q}^{\boldsymbol{i}} \rangle + \langle \boldsymbol{k}^{i-1}, F(\boldsymbol{Q}) - F(\boldsymbol{Q}^{\boldsymbol{i}}) \rangle \geq 0$

$\langle \boldsymbol{U}^i, \boldsymbol{Q} - \boldsymbol{Q}^* \rangle + \langle \boldsymbol{k}^*, F(\boldsymbol{Q}) - F(\boldsymbol{Q}^*) \rangle \geq 0$

for some $\boldsymbol{U}^i \in \partial f_\gamma(\boldsymbol{Q}^{\boldsymbol{i}})$ and some $\boldsymbol{U}^* \in f_\gamma(\boldsymbol{Q}^*)$. As these two inequalities are nearly the same, we only need to prove one of them. For the first inequality, $\boldsymbol{Q}^{\boldsymbol{i}}$ minimizes $L(\boldsymbol{Q}, \boldsymbol{k}^{i-1})$ over all $\boldsymbol{Q}$ and therefore, $\boldsymbol{U}^i \in \partial f_\gamma(\boldsymbol{Q}^{\boldsymbol{i}})$ and $\boldsymbol{U}_j^i \in \partial f_\gamma(\boldsymbol{Q}^{\boldsymbol{i}})$, $1 \leq j \leq m$, such that:

$\boldsymbol{U}^i+\sum_{j=1}^{m} k_j^{i-1}\,\boldsymbol{U}_j^i=0$.

Therefore, the following equation holds

$$\langle \boldsymbol{U}^i, \boldsymbol{Q}-\boldsymbol{Q}^{\boldsymbol{i}}\rangle + \sum_{j=1}^{m} k_j^{i-1}(f_j(\boldsymbol{Q})-f_j(\boldsymbol{Q}^{\boldsymbol{i}})) \geq \langle \boldsymbol{U}^i + \sum_{j=1}^{m} k_j^{i-1}\,\boldsymbol{U}_j^i, \boldsymbol{Q}-\boldsymbol{Q}^{\boldsymbol{i}}\rangle = 0$$

Now, sum the two inequalities and get

$$\langle \boldsymbol{U}^i-\boldsymbol{U}^*, \boldsymbol{Q}^i-\boldsymbol{Q}^*\rangle + \langle \boldsymbol{K}^{i-1}-\boldsymbol{k}^*, F(\boldsymbol{Q}^{\boldsymbol{i}})-F(\boldsymbol{Q}^*)\rangle \leq 0$$

Inspired by the proof for $\|F(\boldsymbol{Q})-F(\boldsymbol{K})\| \leq L(F)\|\boldsymbol{Q}-\boldsymbol{K}\|_F$. The next step is following that

$$\langle \boldsymbol{k}^i-\boldsymbol{k}^*, F(\boldsymbol{Q}^{\boldsymbol{i}})-F(\boldsymbol{Q}^*)\rangle \leq -\langle \boldsymbol{U}^i-\boldsymbol{U}^*, \boldsymbol{Q}^i-\boldsymbol{Q}^*\rangle \leq -\left\|\boldsymbol{Q}^i-\boldsymbol{Q}^*\right\|_F^2$$

We already know that $\boldsymbol{k}^* = [\boldsymbol{k}^* + \boldsymbol{\varphi_i} F(\boldsymbol{Q})]_+$, we can have

$$\left\|\boldsymbol{k}^i-\boldsymbol{k}^*\right\| = \left\|\left[\boldsymbol{k}^{i-1}+\varphi_i F\left(\boldsymbol{Q}^i\right)\right]_+ - [\boldsymbol{k}^*+\varphi_i F(\boldsymbol{Q}^*)]_+\right\|$$

$$\leq \left\|\boldsymbol{k}^{i-1}-\boldsymbol{k}^*+\varphi_i(F(\boldsymbol{Q}^{\boldsymbol{i}})-F(\boldsymbol{Q}^*))\right\|$$

Any projections onto a convex set $\mathbb{R}_+^m$ is a contracting projection. Therefore,

$$\left\|\boldsymbol{k}^i-\boldsymbol{k}^*\right\|^2 = \left\|\boldsymbol{k}^{i-1}-\boldsymbol{k}^*\right\|^2 + 2\varphi_i\langle \boldsymbol{k}^{i-1}-\boldsymbol{k}^*, F(\boldsymbol{Q}^{\boldsymbol{i}})-F(\boldsymbol{Q}^*)\rangle + \varphi_i^2\left\|F\left(\boldsymbol{Q}^{\boldsymbol{k}}\right)-F(\boldsymbol{Q}^*)\right\|^2$$

$$\leq \left\|\boldsymbol{k}^{i-1}-\boldsymbol{k}^*\right\|^2 - 2\varphi_i\left\|\boldsymbol{Q}^i-\boldsymbol{Q}^*\right\|_{\boldsymbol{F}}^{\boldsymbol{2}} + \varphi_i^2 L^2\left\|\boldsymbol{Q}^i-\boldsymbol{Q}^*\right\|_{\boldsymbol{F}}^{\boldsymbol{2}}$$

We replace $L(F)$ by L for short. By assuming the size of φ_i, we have $2\varphi_i - \varphi_i^2 L^2 \geq \beta$ for all $k \geq 1$ and some $\beta > 0$. Finally, we get

$\left\|\boldsymbol{k}^i-\boldsymbol{k}^*\right\|^2 \leq \left\|\boldsymbol{k}^{i-1}-\boldsymbol{k}^*\right\|^2 - \beta\left\|\boldsymbol{Q}^i-\boldsymbol{Q}^*\right\|_{\boldsymbol{F}}^{\boldsymbol{2}}$, and the conclusion as stated in Lemma 3.

Proof for Theorem 3 (Cai et al., 2008):

Set $F_0 = \frac{1}{2}\|\boldsymbol{Q}-\boldsymbol{K}\|_F^2 + \gamma\|\boldsymbol{Q}\|_*$, which is strictly convex. As we can get a unique solution matrix, we define a subgradient of convex function f to express transition from number sets with different dimensions. If we define $\boldsymbol{Z} \in \partial f(\boldsymbol{Q}_0)$, then

$f(\boldsymbol{Q}) \geq f(\boldsymbol{Q}_0) + \langle \boldsymbol{Z}, \boldsymbol{Q} - \boldsymbol{Q}_0 \rangle$. Denote $\widehat{\boldsymbol{K}}$ as the minimizer of F_0 if and only if F_0 contains matrix of all zeros as one subgradient at point $\widehat{\boldsymbol{K}}$. We can derive:

$\gamma \partial \left\| \widehat{\boldsymbol{K}} \right\|_* + \widehat{\boldsymbol{Q}} - \boldsymbol{K} \ni \boldsymbol{0}$, where $\partial \left\| \widehat{\boldsymbol{K}} \right\|_*$ is the subgradient of nuclear norm. Then, $\partial \left\| \widehat{\boldsymbol{K}} \right\|_*$ can be expressed as:

$\partial \|\boldsymbol{K}\|_* = \{\boldsymbol{V}\boldsymbol{U}^* + \boldsymbol{W}\}$, where $\boldsymbol{V}^*\boldsymbol{W} = 0$, $\boldsymbol{W}\boldsymbol{U} = 0$ and $\|\boldsymbol{W}\|_2 \leq 1$.

Apply SVD to decompose $\boldsymbol{Q}$ as $\boldsymbol{Q} = \boldsymbol{V}_0 \boldsymbol{\Sigma}_0 \boldsymbol{U}_0^* + \boldsymbol{V}_1 \boldsymbol{\Sigma}_1 \boldsymbol{U}_1^*$, where $\boldsymbol{V}_0$, $\boldsymbol{U}_0$, $\boldsymbol{V}_1$ and $\boldsymbol{U}_1$ are singular matrices with singular values larger than γ. Based on these derivation, we have $\widehat{\boldsymbol{K}} = \boldsymbol{V}_0(\boldsymbol{\Sigma}_0 - \gamma \boldsymbol{I})\boldsymbol{U}_0^*$, $\boldsymbol{Q} - \widehat{\boldsymbol{K}} = \gamma(\boldsymbol{V}_0 \boldsymbol{U}_0^* + \boldsymbol{W})$ and $\boldsymbol{W} = \gamma^{-1} \boldsymbol{V}_1 \boldsymbol{\Sigma}_1 \boldsymbol{U}_1^*$. Since diagonal elements in $\boldsymbol{\Sigma}_1$ are all less than γ, we can conclude $\gamma \partial \left\| \widehat{\boldsymbol{K}} \right\|_* \ni \widehat{\boldsymbol{Q}} - \boldsymbol{K}$ to finish proof for Theorem 3.

The proof work for other theorems and lemmas are close to the proof for previous lemmas and theorems. All detailed proof work for all lemmas and theorems are done by Candès, E. J., and B. Recht (2009) and Cai et al. (2008)