Nonnegative Matrix Completion for Life-cycle Assessment and Input-ouput Analysis

Fangfang Xu[1], Chen Lin[*,2], Guoping He[3], Zaiwen Wen[4]

Abstract

Life-cycle assessment (LCA) and input-output analysis (IOA) are two data-intensive approaches whose reliability and applicability are dependent on the quality of the data. The difficulty is that not all data is available due to technical or cost reasons. However, the processes producing similar commodities have similar data structures in LCA, while the sectors of the historical surveyed years in IOA have similar input structures as the sectors of the objective year. These features imply that the data usually has low-rank or approximately low-rank structures which enables us to apply the emerging techniques of low-rank matrix completion to recover the missing data. Since the data should be nonnegative in LCA and IOA if we ignore the minor by-product issue, we propose two models for nonnegative matrix completion to recover the missing information of LCA and IOA. The alternating direction method of multipliers is then applied to solve them. The applicability and

[*]Corresponding author. Tel.:+86 13589112903; fax: +86 0531 88364981.
Email address: `c_lin@sdu.edu.cn` (Chen Lin)
[1]Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, China
[2]The Center for Economic Research, Shandong University, Jinan 250100, China
[3]College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao 266510, China
[4]Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China

efficiency of our methods are demonstrated in an application to the widely used database “Ecoinvent” for the life-cycle assessment. Recovering results show that our approaches are helpful.
Key words: Missing data recovery, Nonnegative matrix completion, Life-cycle assessment, Input-output analysis, Alternating direction method of multipliers

---

1. Introduction

Matrix completion (MC) is the process of recovering the unknown or missing elements of a matrix. Under some reasonable assumptions, an incomplete but low-rank matrix can be reconstructed exactly (Candès and Recht, 2009; Recht et al., 2010). The matrix completion problem is widely applicable in many fields, such as machine learning(Abernethy et al., 2006), control(Mesbahi and Papavassilopoulos, 1997), image and video processing(Osher et al., 2005; Ji et al., 2010; Candès et al., 2011), where matrices with low-rank structures are used in the model construction. Another potentially applicable field is the life-cycle assessment (LCA). LCA is a methodology to assess the impacts associated with all the stages of a product's life from-cradle-to-grave, which is sensitive to the data quality. A technology matrix consisting of production/consumption processes can be used to obtain the result of life-cycle inventory (LCI) (Heijungs and Suh, 2002). The technology matrix for LCI can suffer from the missing data problem for both technical and cost reasons. For example, some types of inputs and outputs are difficult to be attributed to a co-product from the multi-output process. The reliability and thus the applicability of the results of an LCI is dependent on the quality of the original data providing the background for the assessment (Weidema and Wesnæs, 1996). The technology matrix of LCI is usually low-rank due to the existence of enormous similar processes. For example, the processes producing similar commodities and the analogous processes in different countries have similar data structures (Huele and van den Berg, 1998). Thus, the matrix completion method can be used to recover the missing data.

IOA is another field of potential application of matrix completion, which has similar data structure with LCA (Heijungs and Suh, 2002; Suh, 2004). In the countries who have the input-output (IO) tables, the statistic departments publish the IO tables only for certain years by survey (Miller and Blair, 2009). For example, China, Japan, and US publish the survey tables for every 5 years (National Bureau of Statistics of China, 2012; Minstry of Internal Affairs and Communication of Japan, 2012; U.S. Bureau of Economic Analysis, 2012). However, when the economists want to do the economic analysis for the other non-surveyed years, data for the objective year needs to be estimated. Only a limited statistical information can be used to estimate the input-output table of the non-surveyed years, while the part of the table without statistical information are missing. The statistic information and historical IO tables, whose sectors have high correlation with the objective year's sectors, can be used to estimate the IO tables of the objective year by using the matrix completion method. It is because the correlativity makes the matrix in question be low-rank or approximately low-rank. In the context of IO, our method can be an alternative estimation method of non-surveyed IO tables other than RAS.

Additionally, Suh (2004) proposed the hybrid LCA method by combining the general IO data with the process-based data to extend the system boundary of LCA. The correlation between IO sectors and processes makes the technology matrix used in the hybrid method be low-rank or approximately low-rank. This implies the possibility of applying the matrix completion method to recover the missing data and improve the data quality for hybrid LCA studies, such as Nakamura and Kondo (2006) and Lin (2011). It is important to note that the above mentioned technology matrices in LCA and IO are nonnegative if we ignore the minor by-product issue. This motivates the development of the nonnegative matrix completion.

Recently, there have been extensive researches on low-rank matrix completion (LRMC), which aims to find a matrix of minimum rank under certain linear constraints. However, the model is NP-hard (non-deterministic polynomial-time hard, Natarajan (1995)) due to the combinatorial nature of the rank function. Candès and Recht (2009) and Recht et al. (2010) show that rank function can be replaced with nuclear norm under some reasonable conditions in order to get a convex problem. As we all known, convex problems can be much easily solved and their local optimal solutions are global. More details about LRMC can be found in (Fazel, 2002; Ma et al., 2009; Candès and Tao, 2010; Cai et al., 2010; Wen et al., 2012; Xu et al., 2012) and references therein. From the numerical experiments of the above literatures, recovery results are very good even when matrices are only approximately low-rank. In practice, certain constraint is usually imposed on the recovered matrix, such as positive semi-definite (Chen et al. (2012)) or nonnegative (Xu et al. (2012)) constraints. Therefore, efficient algorithms are needed to conduct these special matrix completion tasks.

In this paper, we aim to develop efficient algorithms of nonnegative matrix completion to recover the missing data in LCA and IOA. Suppose that technology matrix of LCA and IOA is nonnegative and low-rank or approximately low-rank. Therefore, technology matrix can be completed using the solvers of nonnegative matrix completion when some elements are missing. We propose two models based on whether the known elements of the matrix are contaminated by noise or not. The structure of the two models suggests that the alternating direction method of multipliers (ADMM) is suitable to solve them, that is, one can update each of the variables efficiently while fixing the others. Two ADMM-based algorithms are proposed to conduct nonnegative matrix completion. As an application, we apply the proposed algorithms to recover the potentially missing data in Ecoinvent database—a widely used database of LCA. Recovering results show that our new algorithms can be helpful.

The rest of this paper is organized as follows. The applicability of matrix completion in LCA and IOA is discussed in section 2. We present our nonnegative matrix completion models and algorithms in section 3. The numerical experiments on the Ecoinvent database are present in section 4.

The following notations will be used throughout this paper. Upper (lower) face letters are used for matrices (column vectors). All vectors are column vectors, the superscript $(\cdot)^T$ denotes matrix and vector transposition. $Diag(x)$ denotes a diagonal matrix with x on its main diagonal. $\mathbf{0}$ is a matrix of all zeros of proper dimension, I_n stands for the $n \times n$ identity matrix. The trace of $A \in \mathbb{R}^{m\times n}$, i.e., the sum of the diagonal elements of A, is denoted by

$\mathrm{tr}(A)$. The Frobenius norm of $A \in \mathbb{R}^{m\times n}$ is defined as $\|A\|_F = \sqrt{\sum_{i,j} |A_{i,j}|^2}$. The Euclidean inner product between two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ is defined as $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$. The Euclidean inner product between two matrices $A \in \mathbb{R}^{m\times n}$ and $Z \in \mathbb{R}^{m\times n}$ is defined as $\langle A, Z\rangle = \sum_{i,j}(A_{i,j}Z_{i,j}) = \mathrm{tr}(A^\top Z)$. The inequality $A \geq \mathbf{0}$ is element-wise, which means $A_{ij} \geq 0$ for all entries (i, j). Likewise, the equality $A = Z$ means $A_{ij} = Z_{ij}$ for all entries (i, j).

2. Applicability of MC in LCA and IOA

2.1. LCA and hybrid LCA

LCA is a technique to assess the environmental impacts of a product throughout its life cycle. According to Heijungs and Suh (2002) and Suh (2004), the LCA can be calculated by using the calculation structure that is the same with the environmentally extended input-output model (Leontief, 1970; Duchin, 1990),

$$q = B(I - A)^{-1}y, \tag{1}$$

where y refers to the bundle of products of interest, whose ith element refers to the amount of product produced by process i. The matrix A refers to the technology matrix whose element $A_{i,j}$ shows the amount of input from process i in producing a unit of output by process j. Note that if we rule out the minor by-product issues, the matrix A is nonnegative. The matrix I is the identity, B is the environmental intervention, in which element $B_{i,j}$ denotes the amount of environmental intervention i generated in producing a unit of output by process j, and q refers to the total environmental impact.

The A matrix usually has low rank or approximately low rank due to the existence of enormous similar processes. For example, in Ecoinvent V2.1

(Frischknecht et al., 2005), among all 3963 unit-processes there are 228 processes of electricity generation, 55 processes of road vehicle operations, etc. If there are missing data in the original technology matrix A^0 provided by a database—namely a part of elements of the matrix A^0 is unknown while the other part is known, MC can be used to recover the missing data. In the application part of the paper, our approach will be applied to the Ecoinvent's technology matrix.

2.2. IOA

In IOA, the intermediate input coefficient matrix, which is similar to the technology matrix in LCA, is used in the Leontief quantity model and the price model (Leontief, 1936, 1985). As mentioned previously, in the context of IOA, we are interested in estimating the intermediate input matrix of the non-surveyed years. First, by using the limited statistic information, we can make an uncompleted intermediate input coefficient matrix of the objective non-surveyed year, A_t, whose ijth element shows the amount of input from sector i in producing a unit of output by sector j. A part of A_t^0's elements is known by using the statistic information, while the other part is unknown and needs to be recovered. Accomplishing with the known intermediate coefficient input matrices of the surveyed years, we can build an original matrix for MC,

$$A^0 = \begin{pmatrix} A^0_{t-n} & \ldots & A^0_{t-1} & A^0_t & A^0_{t+1} & \ldots & A^0_{t+n} \end{pmatrix}, \tag{2}$$

where $A_i^0,\ i \neq t$, refers to the known intermediate input coefficient matrices of the surveyed year i. We use the original A^0 to estimate the completed

A. The A matrix is nonnegative if we do not consider the minor by-product issues. Because the sectors of the surveyed years are correlated with the sectors of the objective year, A should satisfy the low-rank or approximately low-rank assumption and the MC is applicable. Based on the original A^0 matrix, our MC approach gives the completed

$$A = \begin{pmatrix} A_{t-n} & \dots & A_{t-1} & A_t & A_{t+1} & \dots & A_{t+n} \end{pmatrix}, \tag{3}$$

where A_t is the estimated intermediate input coefficient matrix of the objective year.

3. Nonnegative Matrix Completion

We now propose two models of nonnegative matrix completion to recover the missing information of the nonnegative matrix A mentioned in Section 2.1 and 2.2. Both models can be solved by the alternating direction method of multipliers. For more details about ADMM, please refer to He et al. (2010); Wen et al. (2010); Boyd et al. (2011).

3.1. Models for Nonnegative Matrix Completion

The matrix completion problem with nonnegative constraints is in the following form:

$$\begin{aligned} &\min \operatorname{rank}(A) \\ &\text{s.t. } A_{ij} = A^0_{ij}, \quad \text{for all } (i,j) \in \Omega, \\ &\qquad A \geq \mathbf{0}, \end{aligned} \tag{4}$$

where $A \in \mathbb{R}^{m \times n}$ is the decision variable, and $A^0_{i,j} \in \mathbb{R}$ are given known elements of matrix A for $(i,j) \in \Omega \subset \{(i,j) : 1 \leq i \leq m, 1 \leq j \leq n\}$. Let

$\mathcal{P}_\Omega$ be the projection onto the subspace of matrices with non-zeros elements restricted to the index subset Ω, i.e.,

$$\mathcal{P}_\Omega(A)_{ij} = \begin{cases} A_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{otherwise.} \end{cases}$$

It follows from the definition of $\mathcal{P}_\Omega$ that the equality constraints in Model 4 can be reformulated as $\mathcal{P}_\Omega(A) = \mathcal{P}_\Omega(A^0)$. Due to the combinational property of the objective function rank$(\cdot)$, Model 4 is NP-hard in general. Inspired by the success of matrix completion under nuclear norm in Candès and Recht (2009); Recht et al. (2010); and Candès and Tao (2010), we use the nuclear norm as an approximation to rank(A) to estimate the optimal solution A^* of Model 4 from the nuclear norm minimization problem with linear and nonnegative constraints:

$$\begin{aligned} \min\ & \|A\|_* \\ \text{s.t. } & \mathcal{P}_\Omega(A) = \mathcal{P}_\Omega(A^0), \\ & A \geq \mathbf{0}, \end{aligned} \tag{5}$$

where the nuclear norm $\|A\|_*$ of A is defined as the summation of the singular values of A, i.e.,

$$\|A\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(A),$$

where $\sigma_i(A)$ is the ith largest singular value of A.

If the known elements of the matrix A are exact, that is to say $\mathcal{P}_\Omega(A^0)$ is reliable, Model 5 is suitable to treat this noise free case. On the contrary, the constraint $\mathcal{P}_\Omega(A) = \mathcal{P}_\Omega(A^0)$ must be relaxed if $\mathcal{P}_\Omega(A^0)$ is contaminated by noise, resulting in either problem:

$$\min_{A \in \mathbb{R}^{m \times n}} \ \|A\|_*, \quad \text{s.t.} \ ||\mathcal{P}_\Omega(A) - \mathcal{P}_\Omega(A^0)|| \leqslant \delta, \ A \geq \mathbf{0}, \tag{6}$$

or the nuclear norm regularized linear least squares model with nonnegative constraints:

$$\min_{A \in \mathbb{R}^{m \times n}} \ \mu\|A\|_* + \frac{1}{2}\|\mathcal{P}_\Omega(A) - \mathcal{P}_\Omega(A^0)\|_2^2, \quad \text{s.t.} \ A \geq \mathbf{0}. \tag{7}$$

Here, δ and μ are given parameters, whose values should be set according to the noise level. Models 6 and 7 are equivalent when the parameters μ and δ are set properly. In this paper, we choose Model 7 to treat the condition in which the known elements of the matrix of interest are contaminated by noise, but our approach can be extended to Model 6 without any difficulty.

Model 7 is especially useful in the context of LCA and IOA. The reason is that the known elements of the technology matrices are usually gotten from large surveys and contaminated by sampling error inevitably. For the sources of uncertainty and variability, we refer the reader to Table 1 in Lloyd and Ries (2007).

3.2. An Alternating Direction Method of Multipliers for Model 5

In this subsection, we present an ADMM-based algorithm for Model 5. To facilitate an efficient use of ADMM, we introduce a new matrix variable

Z and consider an equivalent form of Model 5:

$$\begin{aligned}\min_{A,Z}\ & \|A\|_* \\ \text{s.t.}\ & \mathcal{P}_\Omega(Z) = \mathcal{P}_\Omega(A^0), \\ & A = Z, \\ & Z \geq \mathbf{0},\end{aligned} \tag{8}$$

where $A, Z \in \mathbb{R}^{m\times n}$. The augmented Lagrangian function of Model 8 is:

$$\begin{aligned}\mathcal{L}(A, Z, \Pi, \Lambda) = \|A\|_* + \langle \Pi, \mathcal{P}_\Omega(Z) - \mathcal{P}_\Omega(A^0)\rangle + \langle \Lambda, A - Z\rangle \\ + \frac{\alpha}{2}\|\mathcal{P}_\Omega(Z) - \mathcal{P}_\Omega(A^0)\|_F^2 + \frac{\beta}{2}\|A - Z\|_F^2,\end{aligned} \tag{9}$$

where Π, $\Lambda \in \mathbb{R}^{m\times n}$ are Lagrangian multipliers, $\alpha, \beta > 0$ are penalty parameters.

The alternating direction method of multipliers for (5) is derived by successively minimizing $\mathcal{L}$ with respect to A, Z, Π, and Λ, one at a time while fixing the others at their most recent values, i.e.,

$$A_{k+1} := \arg\min\ \mathcal{L}(A, Z_k, \Pi_k, \Lambda_k), \tag{10a}$$

$$Z_{k+1} := \arg\min_{Z\geq\mathbf{0}}\ \mathcal{L}(A_{k+1}, Z, \Pi_k, \Lambda_k), \tag{10b}$$

$$\Pi_{k+1} := \Pi_k + \gamma\alpha(\mathcal{P}_\Omega(Z_{k+1}) - \mathcal{P}_\Omega(A^0)), \tag{10c}$$

$$\Lambda_{k+1} := \Lambda_k + \gamma\beta(A_{k+1} - Z_{k+1}), \tag{10d}$$

where $\gamma \in (0, 1.618)$. By rearranging the terms of (10a), it is easy to show that it is equivalent to

$$\min_{A \in \mathbb{R}^{m \times n}} \ \frac{1}{\beta}\|A\|_* + \frac{1}{2}\|A - (Z_k - \frac{1}{\beta}\Lambda_k)\|_F^2. \tag{11}$$

Lemma 3.1. *(Theorem 3 in Ma et al. (2009)) Given a matrix $Y \in \mathbb{R}^{m \times n}$ with* $\mathrm{rank}(Y) = t$*, let its Singular Value Decomposition (SVD) be $Y = U_Y \mathrm{diag}(w) V_Y^T$, where $U_Y \in \mathbb{R}^{m \times t}$, $w \in \mathbb{R}_+^t$, $V_Y \in \mathbb{R}^{n \times t}$, and $\nu \geq 0$. Define the shrinkage operator $s_\nu(\cdot)$ as*

$$s_\nu(w) = \bar{w}, \ \textit{with } \bar{w}_i = \begin{cases} w_i - \nu, & \textit{if } w_i - \nu > 0, \\ 0, & \textit{otherwise.} \end{cases}$$

Then

$$A := \mathcal{S}_\nu(Y) = U_Y Diag(s_\nu(w)) V_Y^T$$

is an optimal solution of the problem:

$$\min_{A \in \mathbb{R}^{m \times n}} \ f(A) := \nu\|A\|_* + \frac{1}{2}\|A - Y\|_F^2.$$

Based on Lemma 3.1, we can obtain the solution of Model 11:

$$A_{k+1} = \mathcal{S}_{\frac{1}{\beta}}(Z_k - \frac{1}{\beta}\Lambda_k).$$

Model 10b is equivalent to

$$\begin{aligned}\min_{Z\in\mathbb{R}^{m\times n}}\ & \langle \Pi_k, \mathcal{P}_\Omega(Z) - \mathcal{P}_\Omega(A^0)\rangle - \langle \Lambda_k, Z\rangle \\ & + \frac{\alpha}{2}\|\mathcal{P}_\Omega(Z) - \mathcal{P}_\Omega(A^0)\|_F^2 + \frac{\beta}{2}\|Z - A_{k+1}\|_F^2 \\ \text{s.t.}\ \ & Z \geq \mathbf{0}.\end{aligned} \tag{12}$$

We replace the quadratic term $\|\mathcal{P}_\Omega(Z) - \mathcal{P}_\Omega(A^0)\|_2^2$ in (12) by its first-order Taylor expansion with a proximal term $\frac{\rho}{2}\|Z - Z_k\|_F^2$ and obtain an approximate model:

$$\begin{aligned}\min\ & \langle \Pi_k, \mathcal{P}_\Omega(Z) - \mathcal{P}_\Omega(A^0)\rangle + \alpha\langle \mathcal{P}_\Omega(Z_k) - \mathcal{P}_\Omega(A^0), Z - Z_k\rangle \\ & - \langle \Lambda_k, Z\rangle + \frac{\rho\alpha}{4}\|Z - Z_k\|_F^2 + \frac{\beta}{2}\|Z - A_{k+1}\|_F^2 \\ \text{s.t.}\ \ & Z \geq \mathbf{0}.\end{aligned} \tag{13}$$

By rearranging the terms of the objective function of Model 13, it is easily verified that it is equivalent to

$$Z_{k+1} := \arg\min_{Z\in\mathbb{R}_+^{m\times n}}\ \|Z - G_k\|_F^2,$$

whose solution is

$$Z_{k+1} = \mathcal{Q}_+(G_k),$$

where

$$G_k = \frac{2}{\rho\beta + 2\beta}[\frac{\rho\alpha}{2}Z_k + \Lambda_k + \beta A_{k+1} - \mathcal{P}_\Omega(\Pi_k) - \alpha(\mathcal{P}_\Omega(Z_k) - \mathcal{P}_\Omega(A^0))],$$

$$(\mathcal{Q}_+(G_k))_{ij} = \max\{(G_k)_{ij}, 0\}.$$

In short, ADMM with linearization applied to Model 5 produces the iteration:

$$\begin{aligned} A_{k+1} &:= \mathcal{S}_{\frac{1}{\beta}}(Z_k - \frac{1}{\beta}\Lambda_k), && \text{(14a)} \\ Z_{k+1} &:= \mathcal{Q}_+(G_k), && \text{(14b)} \\ \Pi_{k+1} &:= \Pi_k + \gamma\alpha(\mathcal{P}_\Omega(Z_{k+1}) - \mathcal{P}_\Omega(A^0)), && \text{(14c)} \\ \Lambda_{k+1} &:= \Lambda_k + \gamma\beta(A_{k+1} - Z_{k+1}). && \text{(14d)} \end{aligned}$$

Algorithm 1 below shows a pseudo code for the proposed algorithm for Model 5 in which known elements of the matrix of interest do not suffer from noise.

Algorithm 1: ADMM-based method for noise-free nonnegative matrix completion

1 Input $\mathcal{P}_\Omega(A^0)$, *maxiter* ≥ 0, and *tol* ≥ 0.
2 Set γ, α, β, and $\rho \geq 0$. Set Z_0 as a nonnegative random matrix, and Π_0, Λ_0 as zero matrices of appropriate sizes.
3 **while** *not converge* **do**
4 Update $A_k, Z_k, \Pi_k, \Lambda_k$ by the formulas (14).

Similarly, ADMM applied to Model 7 generates the iteration (15) and Algorithm 2.

$$\begin{aligned} Z_{k+1} &:= \mathcal{Q}_+(Z_k + \frac{1}{\rho}\Lambda_k), && \text{(15a)} \\ A_{k+1} &:= S_{\frac{\mu}{\rho+\beta}}(G_k), && \text{(15b)} \\ \Lambda_{k+1} &:= \Lambda_k + \gamma\rho(A_{k+1} - Z_{k+1}), && \text{(15c)} \end{aligned}$$

where $G_k = \frac{1}{\rho+\beta}(\beta A_k + \rho Z_{k+1} - (\mathcal{P}_\Omega(A_k) - \mathcal{P}_\Omega(A^0)) - \Lambda_k)$.

Algorithm 2: ADMM-based method for nonnegative matrix completion with noise

Input $\mathcal{P}_\Omega(A^0)$, $maxiter \geq 0$, and $tol \geq 0$.
Set ρ, μ, γ, and $\beta \geq 0$. Set Z_0 as a random matrix, and Λ_0 as zero matrices of appropriate sizes.
while *not converge* **do**
  Update Z_k, A_k, Λ_k by the formulas (15).

The convergence of the above two schemes is ensured by the theory in (Hong and Luo, 2012).

4. Application

In this section, Algorithms 1 and 2 are applied to the recovery of the potentially missing data in Ecoinvent database V2.1 (Frischknecht et al., 2005). There are 3964 unit-processes to build the original technology matrix A^0 for recovery, which is a 3964×3964 square matrix. Its element $A^0_{i,j}$ shows the amount of input from process i in producing a unit of output by process j. We assume that there are missing data in A^0 matrix given by Ecoinvent database and our goal is to complete the matrix by using our method. The non-zero elements in A^0 are considered to be known elements. However, we can not argue that all the zero elements in A^0 are potentially unknown elements. It is because not all the 3964×3964 elements in the matrix should have non-zero values, since technically there may not be any input from a certain process to another certain process. In order to find the should-be-zero elements, we relate every Ecoinvent process to an sector of the US input-output (IO) table. Multiple Ecoinvent processes can belong to an IO sector because the IO sector is more general than Ecoinvent processes.

To be exact, let the IO sector related to Ecoinvent process i be sector i_o. Let Θ be the set of the should-be-zero elements. Then we define $A^0_{i,j} \in \Theta$ if $B_{i_o j_o} = 0$, where B_{i_o,j_o} refers to the element of the intermediate input matrix B of US IO table. This means that if there exists no input from IO sector i_o to IO sector j_o, there should be no input from the Ecoinvent processes in IO sector i_o to the Ecoinvent processes belonging to IO sector j_o. In this case, the size of Θ is 3,197,414. Let the set of non-zero element in A^0 be Γ (31,950 elements). Consequently, the set of known elements can be given by $\Omega = \Gamma \cup \Theta$, which has 3,229,364 elements. The information from Ω is used to recover the potentially missing data in A^0. We rule out the minor by-product issues, which implies that the A^0 matrix is nonnegative. With regard to the environmental pressure data (or output flow data) of Ecoinvent, although its recovery is not included in the current application, our approach is also applicable because of the similarity of data structure.

All algorithms are implemented in MATLAB and all experiments were performed on a Dell Precision T5500 workstation with Intel(R) Xenon(R) E5620 CPU at 2.40GHz ($\times 4$) and 12GB of memory running Ubuntu 12.04 and MATLAB 2011b.

4.1. Experiments on noise-free Data

As mentioned previously, 3,229,364 elements (about 20% of elements in A^0) are known, the rest (12,483,932 elements) are unknown or missing. In order to check the accuracy of our algorithms, we select a subset of a% uniformly at random from the known elements set Ω and denote it by Ω_1. The residual $(100-a)$% of the known elements is denoted by Ω_2. We use Ω_1 to conduct nonnegative matrix completion, then use Ω_2 to compute precision of our algorithms. In summary, we divide the known elements set Ω to Ω_1 (used for nonnegative matrix completion) and Ω_2 (used for computing precision).

For Algorithm 1, we set $\gamma = 1.618$, the value of $\alpha = 0.1$, $\beta = 0.1$, and $\rho = 0.01$. The algorithm is stopped if a maximal number of 10^3 iterations is reached or if the following conditions are satisfied:

$$\frac{\|A_{k+1} - A_k\|_F}{\max(1, \|A_k\|_F)} \le tol, \tag{16a}$$

$$\|A_{k+1} - Z_{k+1}\|_F \le tol, \tag{16b}$$

$$\|\mathcal{P}_\Omega(Z_{k+1}) - \mathcal{P}_\Omega(A^0)\|_2 \le tol, \tag{16c}$$

where tol is set to be 10^{-4}. Finally, Algorithm 1 outputs the completed A as the estimation of the full technology matrix based on A^0. The average relative errors of the elements belonging to Ω_1 and Ω_2 are denoted by $err1$ and $err2$, respectively:

$$err1 = \frac{\sqrt{\sum_{(i,j)\in\Omega_1}(\frac{A_{ij}-A^0_{ij}}{\max(1,A^0_{ij})})^2}}{|\Omega_1|},$$

$$err2 = \frac{\sqrt{\sum_{(i,j)\in\Omega_2}(\frac{A_{ij}-A^0_{ij}}{\max(1,A^0_{ij})})^2}}{|\Omega_2|}.$$

Both $err1$ and $err2$ are computed when a is equal to 80, 85, 90, 95, or 100. The results are shown in Table 1, where p is the cardinality of set Ω_1 (number of the elements in Ω_1). p is also the number of the known elements used to conduct nonnegative matrix completion.

From Table 1, we can observe that when a increase from 80 to 100, $err1$

Table 1: Accuracy Results

	Sampling Ratio (SR)				
a%	80 %	85 %	90 %	95 %	100 %
p	2590671	2750344	2910017	3069690	3229364
err1	2.27×10^{-6}	2.14×10^{-6}	2.02×10^{-6}	1.91×10^{-6}	1.82×10^{-6}
err2	8.36×10^{-4}	1.06×10^{-3}	9.69×10^{-4}	1.05×10^{-3}	0
err	8.38×10^{-4}	1.06×10^{-3}	9.71×10^{-4}	1.06×10^{-3}	1.82×10^{-6}

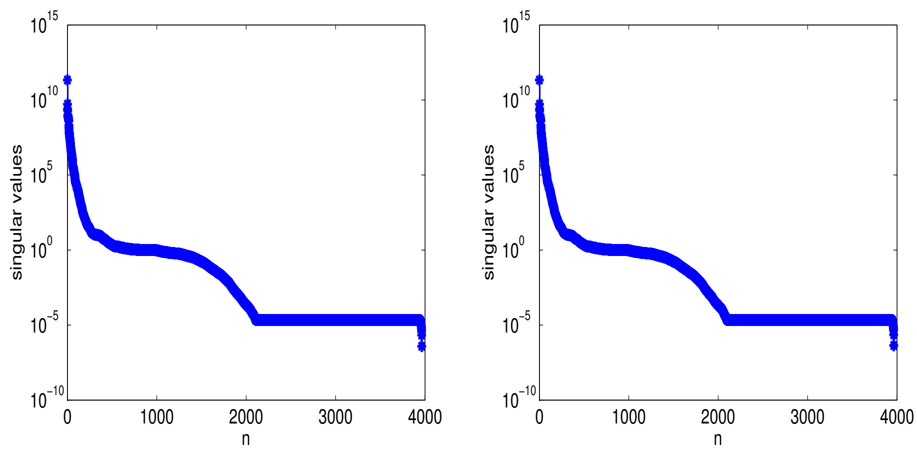

Figure 1: The singular values of the matrix A. The left figure corresponds to $a = 95$ case, and the right one corresponds to $a = 100$ case. the horizontal axis refers to the number of singular value, while the n-th point on the curve denotes the n-th largest singular value of the completed matrix A.

decreases monotonously. The more known elements there are, the better completed matrix will be. When a is equal to 100, the set Ω_2 is empty, which means that $err2$ is equal to 0. All the values of $err2$ are small, it indicates that elements of completed matrix A are closed to their true value. In conclusion, Algorithm 1 outputs satisfactory numerical results when the known elements are noise free.

The singular values of the completed matrix A are depicted in Figure 1.

We observe that the singular values decrease quickly and there are only a few dominant singular values. Specifically, the 500th largest singular value is already close to 1. Therefore, A is low-rank or approximately low-rank.

4.2. Experiments on data with Noise

We now apply Algorithm 1 and 2 to the recovery of Ecoinvent technology matrix when the known elements are contaminated by noise.

The parameter μ in Model 7 is usually set to be a medium value between 0 to 10 in general. We can use continuation technique to speed the convergence of Algorithm 2. We solve a sequence of Model 7 defined by a decreasing sequence $\{\mu^0,\ \mu^1, \ldots,\ \mu^l = \bar{\mu}\}$ with a finite positive integer l. We set $\bar{\mu} = 10^{-4}\mu^0$ with the initial $\mu^0 = 100$, and update $\mu^k = \max\{0.7\mu^{k-1},\ \bar{\mu}\}$ at iteration k. All the other parameters are set to be the same as Section 5.1.

Suppose that $\mathcal{P}_\Omega(A^0)$ is contaminated and ϵ is the error ratio. We select M between $\mathcal{P}_\Omega(A^0) - \epsilon\mathcal{P}_\Omega(A^0)$ and $\mathcal{P}_\Omega(A^0) + \epsilon\mathcal{P}_\Omega(A^0)$ at random, that is to say, $\mathcal{P}_\Omega(A^0) - \epsilon\mathcal{P}_\Omega(A^0) \leq M \leq \mathcal{P}_\Omega(A^0) + \epsilon\mathcal{P}_\Omega(A^0)$. Then M is used to conduct nonnegative matrix completion using Algorithm 1 and 2. The results are shown in Table 2 corresponding to different ϵ. In this subsection, a is set to be 95.

Table 2: Accuracy Results

	$\epsilon = 0.001$		$\epsilon = 0.01$		$\epsilon = 0.1$	
	Algorithm1	Algorithm2	Algorithm1	Algorithm2	Algorithm1	Algorithm2
err1	1.90×10^{-6}	8.42×10^{-6}	1.93×10^{-6}	8.41×10^{-6}	2.64×10^{-6}	8.53×10^{-6}
err2	8.10×10^{-4}	6.32×10^{-4}	9.65×10^{-4}	7.44×10^{-4}	1.30×10^{-3}	9.90×10^{-4}
err	8.12×10^{-4}	6.41×10^{-4}	9.67×10^{-4}	7.53×10^{-4}	1.30×10^{-3}	9.99×10^{-4}

From Table 2, we observe that *err* of Algorithm 2 is smaller than that of Algorithm 1 corresponding to certain ϵ. When ϵ increases from 0.001 to 0.1,

err of Algorithm 2 increases slower than that of Algorithm 1. Therefore, the recovery results of Algorithm 2 are better than those of Algorithm 1 when the known elements are contaminated. It also means that Algorithm 2 is more robust.

4.3. An Application on Greenhouse Gas Assessment

In order to show the effect of the matrix completion on LCA results, we apply the completed Ecoinvent data to the assessment of greenhouse gas impact induced by products' supply chain. Equation (1) is used as the calculation method. In this application, $B = CE$, where E is the environment pressure matrix given by Ecoinvent and C is the characterization factors of GWP100 (global warming potential 100 years) provided by IPCC 2001 (intergovernmental panel on climate change) (McCarthy et al., 2001). We let $y = \iota$, where ι is a vector of all ones of proper dimension. Thus, we assess the summation of the GWP100 impact induced by one unit of each process's product.

We now calculate the GWP impact given by the original technology matrix (A^0) and the completed technology matrix (A when $a = 100$ and data is noise free). Since neither $I - A^0$ nor $I - A$ are invertible, we take 1000 processes from the 3964 processes and build dimension-reduced versions of A^0 and A, which are all invertible. The results are shown in Figure 2. GWP given by the original technology matrix is 1,168 kg CO2-Eq, while that obtained from the completed matrix is 1,176 kg CO2-Eq. In other words, when the completed technology matrix A is applied, the GWP impact increases by 0.63%.

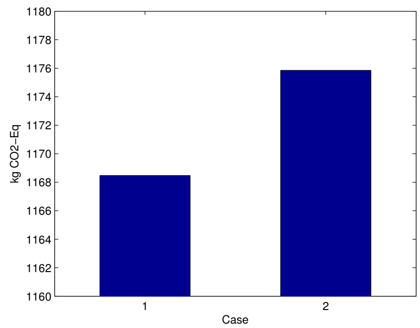

Figure 2: Comparison of GWPs given by different technology matrices. Case 1 refers to GWP in kg CO2-Eq when $q = B(I - A^0)^{-1}y$. Case 2 refers to GWP in kg CO2-Eq when $q = B(I - A)^{-1}y$.

5. Discussion

In this paper, two algorithms are proposed to estimate missing information of LCA and IO data. Firstly, we present two nonnegative matrix completion models based on whether the known elements of the matrix are exact or not. The alternating direction method of multipliers is then applied to solve these models. Recovering results on Ecoinvent database, a widely used database for LCA, show that the new models and algorithms are helpful.

A few directions for future work remains. First, in the context of IOA, a comparison of our method with the conventional RAS method is an important future work. Second, the correlation between IO sectors and processes makes the technology matrix used in the hybrid LCA be low-rank or approximately low-rank. This implies the possibility of applying the our methods to recover the missing data and improve the data quality for hybrid LCA studies. Third, better optimization models and their corresponding algorithms are needed in order to consider other practical constraints required in LCA and IOA.

6. Acknowledgements

The work of F. Xu and G. He was supported in part by NSFC (Grant No. 10971122), Scientific and Technological Projects (2009GG10001012) of Shandong Province and Excellent Young Scientist Foundation (2010BSE06047, BS2012SF025) of Shandong Province. The work of C. Lin was supported in part by Shandong Province Natural Science Foundation (ZR2012GQ004) and Independent Innovation Foundation (IFW12098) of Shandong University. The work of Z. Wen was supported in part by NSFC grant 11101274 and Research Fund (20110073120069) for the Doctoral Program of Higher Education of China.

References

Abernethy, J., Bach, F., Evgeniou, T., et al., 2006. Low-rank matrix factorization with attributes. Tech. rep., arXiv preprint cs/0611124.

Boyd, S., Parikh, N., Chu, E., et al., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning 3, 1–122.

Cai, J.-F., Candés, E. J., Shen, Z., 2010. A singular value thresholding algorithm for matrix completion. SIAM J. Op 20, 1956–1982.

Candès, E. J., Eldar, Y., Strohmer, T., Voroninski, V., 2011. Phase retrieval via matrix completion. Tech. rep., Stanford University.

Candès, E. J., Recht, B., May 2009. Exact matrix completion via convex optimization. Foundations of Computational Mathematics.

Candès, E. J., Tao, T., 2010. The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inform. Theory 56 (5), 2053–2080.

Chen, C., He, B., Yuan, X., 2012. Matrix completion via an alternating direction method. IMA Journal of Numerical Analysis 32, 227–245.

Duchin, F., 1990. The conversion of biological materials and wastes to useful products. Structural Change and Economic Dynamics 1 (2), 243–261.

Fazel, M., 2002. Matrix rank minimization with application. Ph.D. thesis, Stanford University.

Frischknecht, R., Jungbluth, N., Althaus, H., Doka, G., Dones, R., Heck, T., Hellweg, S., Hischier, R., Nemecek, T., Rebitzer, G., et al., 2005. The ecoinvent database: Overview and methodological framework (7 pp). The International Journal of Life Cycle Assessment 10 (1), 3–9.

He, B., Tao, M., Yuan, X., 2010. Alternating direction method with gaussian back substitution for separable convex programming. Tech. rep., Nanjing University.

Heijungs, R., Suh, S., 2002. The computational structure of life cycle assessment. Vol. 11. Kluwer Academic Publishers.

Hong, M., Luo, Z.-Q., Aug. 2012. On the Linear Convergence of the Alternating Direction Method of Multipliers. ArXiv e-prints.

Huele, R., van den Berg, N., 1998. Spy plots. The International Journal of Life Cycle Assessment 3 (2), 114–118.

Ji, H., Liu, C., Shen, Z., et al., 2010. Robust video denoising using low rank matrix completion. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE. pp. 1791–1798.

Leontief, W., 1936. Quantitative input and output relations in the economic systems of the united states. The review of Economics and Statistics 18 (3), 105–125.

Leontief, W., 1970. Environmental repercussions and the economic structure: An input–output approach. Review of Economics and Statistics 52 (3), 262–271.

Leontief, W., 1985. Technological change, prices, wages, and rates of return on capital in the u.s. economy. In: Leontief, W. (Ed.), input–output Economics. Oxford: Oxford University Press, pp. 392–417.

Lin, C., 2011. Identifying lowest-emission choices and environmental pareto frontiers for wastewater treatment input-output model based linear programming. Journal of Industrial Ecology.

Lloyd, S. M., Ries, R., 2007. Characterizing, propagating, and analyzing uncertainty in life-cycle assessment: A survey of quantitative approaches. Journal of Industrial Ecology 11 (1), 161–179.

Ma, S., Goldfarb, D., Chen, L., 2009. Fixed point and bregman iterative methods for matrix rank minimization. Mathematical Programming, 1–33.

McCarthy, J. J., Canziani, O. F., Leary, N. A., Dokken, D. J., White, K. S.,

2001. Climate change 2001: impacts, adaptation, and vulnerability: contribution of Working Group II to the third assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Mesbahi, M., Papavassilopoulos, G. P., 1997. On the rank minimization problem over a positive semidefinite linear matrix inequality. IEEE Transactions on Automatic Control 42, 239–243.

Miller, R., Blair, P., 2009. Input-output analysis: foundations and extensions. Cambridge University Press.

Minstry of Internal Affairs and Communication of Japan, 2012. Input-output table of Japan. Accessed December 2012.
URL http://www.stat.go.jp/data/io/index.htm

Nakamura, S., Kondo, Y., 2006. A waste input–output life-cycle cost analysis of the recycling of end-of-life electrical home appliances. Ecological Economics 57 (3), 494–506.

Natarajan, B. K., 1995. Sparse approximation solutions to linear systems. SIAM J. Comput. 24 (2), 227–234.

National Bureau of Statistics of China, 2012. Input-output table of china. Accessed December 2012.
URL http://www.stats.gov.cn/tjsj/qtsj/trccb/

Osher, S., Burger, M., Goldfarb, D., et al., 2005. An iterative regularization method for total variation-based image restoration. Multiscale Model. Simul. 4, 460–489.

Recht, B., Fazel, M., Parrilo, P. A., 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. 52 (3), 471–501.

Suh, S., 2004. Functions, commodities and environmental impacts in an ecological–economic model. Ecological Economics 48 (4), 451–467.

U.S. Bureau of Economic Analysis, 2012. Benchmark input-output data. Accessed December 2012.
URL `http://www.bea.gov/industry/io_benchmark.htm`

Weidema, B., Wesnæs, M., 1996. Data quality management for life cycle inventoriesan example of using data quality indicators. Journal of Cleaner Production 4 (3), 167–174.

Wen, Z., Goldfarb, D., Yin, W., 2010. Alternating direction augmented lagrangian methods for semidefinite programming. Mathematical Programming Computation 2 (3-4), 203–230.

Wen, Z., Yin, W., Zhang, Y., 2012. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. Math. Prog. Comp. 4, 333–361.

Xu, Y., Yin, W., Wen, Z., Zhang, Y., 2012. An alternating direction algorithm for matrix completion with nonnegative factors. Frontiers of Mathematics in China 7, 365–384.