Multiple Imputation for Large Hierarchical Multidimensional Data with Linear Aggregation Constraints

January 6, 2024

Abstract

The use of multiple imputation of missing data in empirical studies has become increasingly popular in recent years. However, currently available multiple imputation methods face significant challenges when applied to large hierarchical, multidimensional data sets that are subject to linear aggregation constraints. In this paper we introduce a novel multiple imputation method designed to address these challenges. Our method leverages singular multivariate normal distributions within an Expectation Maximization algorithm combined with a Parallel-Sequential Imputation scheme to handle large and complex data sets that include linear aggregation constraints. Testing on real data sets demonstrates that the new method obtains up to twice the accuracy and is as much as an order of magnitude faster than leading alternative methods. We apply our method to estimate a panel data model of average weekly wages and show that our method produces estimates that unbiased and as efficient as estimates based on the dataset with no missing values.

Keywords: missing data, expectation maximization, singular multivariate normal distribution, quasi-Monte Carlo, QCEW

1 Introduction

Missing data arising from subject non-response, system failures, measurement errors, confidentiality constraints or other causes, poses a frequent challenge in empirical economic studies. Because most analytical methods require complete datasets, researchers must often either remove or fill the incomplete observations, or implicitly incorporate the missing data into their models, to facilitate meaningful analyses and inferences. Depending on the mechanism of missing data (Rubin, 1987), appropriate missing data methods are crucial to avoid information loss and biased analyses.

In Economic studies, the most commonly used missing data method is list-wise deletion, which simply excludes incomplete observations. However, this approach causes biased results if the data are not Missing Completely At Random (MCAR). For example, if survey respondents within certain demographic groups are less inclined to report their consumption patterns, analyses based solely on the complete cases would produce biased results. Recently, more sophisticated methods, such as maximum likelihood estimation, weighting, and imputation methods have been developed for scenarios where the MCAR assumption does not hold. Imputation methods have gained increasing popularity since, after imputing the missing values, researchers can freely apply any analytical methods of interest. Multiple imputation (MI) methods, paricularly, have come to dominate because they can generate accurate imputations while also capturing the additional uncertainty of the missing data.

Since the pioneering work of Rubin (1987), a wide array of multiple imputation methods has been developed, including various modeling specifications and sampling techniques to handle different data types. These methods include parametric methods such as Joint Modeling (JM) (Schafer, 1997; Rizopoulos, 2012) and Fully Conditional Specification (FCS) (Azur et al., 2011; Van Buuren, 2018), as well as semi-parametric and non-parametric methods like Hot Deck (Cranmer and Gill, 2013), Predictive Mean Matching (PMM) (Rubin, 1986; Little, 1988), and machine learning-based imputation techniques (Stekhoven and Bühlmann, 2012; Batista et al., 2002). These methods can effectively impute common data types, including continuous, categorical, survival, longitudinal, and panel data (Van Buuren, 2018; Little and Rubin, 2019). Among longitudinal or panel data studies, the most widely used multiple imputation methods include Multivariate Imputation by Chained

Equations (MICE) (Azur et al., 2011), the Expectation Maximization (EM) based methods (Honaker and King, 2010; King et al., 2001), and the Markov Chain Monte Carlo (MCMC) based methods (Gelman et al., 1995; Schafer, 1997). MICE is well known for its flexibility with mixed-type data and varying patterns of missing data. EM-based methods, such as the EM with Bootstrapping (EMB) (Honaker and King, 2010) method, are computationally efficient and effective for large datasets. MCMC-based methods have been shown to be robust for datasets with complex relationships and distributions.

Despite these advances, existing multiple imputation methods struggle to handle large, hierarchical datasets with linear aggregation constraints, i.e. individual values aggregated across dimensions like time, hierarchical levels, or geographic areas. Failure to incorporate these aggregations can lead to constraint violations in imputed data that undermine subsequent analyses. Moreover, neglecting these aggregations risks losing crucial information and potentially causing the Missing Not At Random (MNAR) problem and may significantly reduce imputation quality. Incorporating linear aggregations directly into MICE and EM-based methods is problematic as the aggregations introduce perfect colinearity leading to singular covariance matrices. Extensions of the MCMC-based methods can be specialized to handle linear constraints. Specifically, the constrained Dirichlet process mixture of multivariate normals (CDPMMN) multiple imputation engine (Kim et al., 2014) uses a hit-and-run sampler to ensure the imputed values meet the linear inequality constraints. However, it does not support multidimensional linear aggregation constraints. Also, the Bayesian Multiscale Multiple Imputation (BMMI) method (Holan et al., 2010) uses singular normal distributions to model the linear aggregations into the MCMC process. However,the MCMC process may not be suitable for large datasets because it is computationally intensive and may converge slowly.

In this paper, we introduce a novel method, Multidimensional Bootstrapping Expectation Maximization Multiple Imputation (MBEMMI) method, designed for efficient and accurate imputation of large, hierarchical structured multidimensional data with linear aggregation constraints. MBEMMI uses singular normal distributions to leverage extra information from redundant linear aggregation constraints, thereby enhancing imputation quality and ensuring compliance with these constraints. Additionally, MBEMMI employs an EM algorithm that has deterministic convergence and incorporates a novel Parallel Sequential Imputation (PSI) scheme that allows for the algorithm to parallelized across many processors.

Tests on real data and variants demonstrate that the MBEMMI method is more accurate than the leading MCMC alternative BMMI while maintaining competitive processing speed compared with the EM-based method EMB. Specifically, MBEMMI requires about five minutes to generate ten imputed datasets while EMB requires two minutes and BMMI requires fifty minutes. In an application to estimate a fixed effect model of average weekly wages, MBEMMI yielded unbiased point estimates and recovered standard errors comparable to a complete data scenario, while the complete case (list-wise deletion) study obtained biased estimates and failed to make correct inferences due to large standard errors.

The remainder of the paper is organized as follows: Section 2 discusses the hierarchical, multidimensional data structure with aggregation constraints using a sample of the Quarterly Census of Employment and Wages (QCEW) data. Section 3 describes the MBEMMI method, focusing on its approach to estimate the distribution of missing values. Section 4 adapts MBEMMI, BMMI, and EMB methods to use the Parallel Sequential Imputation scheme for large datasets. Validation of the MBEMMI-PSI algorithm is presented in Section 5. In Section 6 we test our method for model estimation. Finally, Section 7 concludes the paper.

2 Data Structure

Hierarchical multidimensional data structures organize information across several dimensions, each following a hierarchical order. Example are GDP data and QCEW data, available across time and geographic dimensions, with each dimension containing hierarchical levels and higher levels are aggregations of the lower levels.

For illustration, Table 1 shows a sample of the Florida Quarterly Census of Employment and Wage (QCEW) data, as released by the Bureau of Labor Statistics (BLS). This data sample includes five years of employment counts across three sub-industries within the same parent industry. In addition to the quarterly counts, the sample also contains annual aggregations (every fifth row) and industry-wide totals (column 4). To protect the industries that are too small and vulnerable to intruders, BLS suppresses employment and wage data for cells meeting the suppression rule.[1] The suppressed values are indicated by **S** in the Table.

Table 1: Hierarchical multidimensional data example. A subset of the Florida QCEW data including suppressed values marked as **S**.

	Series 1	Series 2	Series 3	Total
year1.q1	20	414	484	918
year1.q2	24	412	493	929
year1.q3	25	404	508	937
year1.q4	23	415	527	965
year1.a	92	1,645	2,012	3,749
year2.q1	9	262	540	811
year2.q2	**S**	**S**	557	839
year2.q3	**S**	**S**	510	831
year2.q4	**S**	**S**	528	868
year2.a	**S**	**S**	2,135	3,349
year3.q1	**S**	**S**	676	1,200
year3.q2	21	495	684	1,200
year3.q3	20	468	665	1,152
year3.q4	**S**	**S**	703	1,217
year3.a	79	1,964	2,728	4,769
year4.q1	32	476	645	1,153
year4.q2	30	473	652	1,155
year4.q3	31	484	686	1,200
year4.q4	30	553	723	1,306
year4.a	123	1,986	2,706	4,814
year5.q1	36	538	630	1,205
year5.q2	41	502	661	1,204
year5.q3	45	500	657	1,202
year5.q4	48	514	639	1,200
year5.a	170	2,054	2,587	4,811

It is challenging to impute the missing quarterly counts, as they are constrained by the annual and industry aggregations. To accurately impute them, multiple imputation methods must incorporate the linear aggregation constraints while estimating the distribution of missing values. Successfully doing so not only makes the imputations meet the constraints,

[1] The BLS does not explicitly disclose its suppression rule, although a widely accepted approximation is the 80/3 rule: a cell number is suppressed if there are fewer than 3 establishments or if any one establishment's employment accounts for more than 80% of the total employment (BLS, 2017).

but also potentially extracts additional information about the missing values from these constraints, thereby enhancing imputation accuracy.

As highlighted in Section 1, popular multiple imputation methods like MICE (Azur et al., 2011) and EMB (Honaker and King, 2010) struggle to accommodate the multidimensional aggregations due to the issue of perfect multicolinearity in the regression models. The BMMI method (Holan et al., 2010) is more capable of handling these aggregations; however, the stochastic convergence of the MCMC process demands considerable expertise to determine whether a convergence has been reached. Moreover, the MCMC process requires an extensive burn-in period to ensure the integrity of the chains and large enough interval to thin the chains and mitigate auto-correlation between consecutive imputations. Consequently, the BMMI method is slow, especially when the dataset is large.

Due to its size, the full QCEW data presents an even greater challenge than the sample previously discussed. Organized using the North American Industry Classification System (NAICS) code, the Florida QCEW dataset contains quarterly employment and wage information across up to 2,678 industries.[2] As detailed in Table 2, within the 2012-2016 Florida QCEW data, 232 out of 2,157 industries are incomplete, with an average missing rate of 59.01% in those incomplete industries.

Table 2: Summary of the 2012-2016 Florida QCEW data. Industry count, incomplete industry count, and mean missing rate of the incomplete industries (95% CI) grouped by NAICS code levels.

Level	Industry Count	Incomplete Count	Incomplete Mean Missing %
2-digit	25	0	-
3-digit	94	2	80% (80%, 80%)
4-digit	316	15	48.67% (34.92%, 62.41%)
5-digit	679	56	60.54% (52.32%, 68.75%)
6-digit	1,043	159	59.18% (54.2%, 64.16%)
Total	2,157	232	59.01% (54.99%, 63.03%)

Imputing large datasets like the QCEW or nationwide consumption surveys, which often contain thousands of units, is extremely time-consuming. The most efficient strategy is parallelization of the imputation processes. However, methods like BMMI require either automatic convergence detection tools or long enough chains to ensure a convergence.

[2] The 2022 version of the NAICS code includes 2,678 industry identifiers.

In addition, priors that meet certain criteria must be provided for each parallel process to guarantee that the imputations obtained in parallel can adequately capture the uncertainty. Although the EMB method can be easily parallelized, it is challenging to account for linear aggregation constraints using this method. As a result, there is currently no multiple imputation method that can accurately and swiftly impute large hierarchical multidimensional data.

3 Multidimensional Bootstrapping Expectation Maximization Multiple Imputation

To describe how the Multidimensional Bootstrapping Expectation Maximization Multiple Imputation (MBEMMI) method works, we will progress through the following steps: modeling the data series, incorporating linear aggregations, and detailing the multiple imputation process.

MBEMMI starts with the assumption that, in the absence of linear aggregations, the data follow a multivariate normal distribution (MVN) in at least one dimension, i.e., $Y \sim \mathcal{N}(\mu, \Sigma)$. The dimension may be time as in the QCEW example shown in Table 1, or geographic areas as in a consumption pattern survey. While the MVN assumption might seem strong, it generally holds true for large datasets like the QCEW. Moreover, researchers have demonstrated that an MVN assumption often performs as well as more complex alternatives in multiple imputation practice (Schafer (1997), Schafer and Olsen (1998)).

Under the MVN assumption, MBEMMI models each variable as a linear function of all other variables. For a missing value $y_{i,j}^{mis}$, the model is represented as:

$$y_{i,j}^{mis} = \mathbf{y}_{i,-j}^{obs}\boldsymbol{\beta} + \epsilon_i \tag{1}$$

where i denotes the observation, j denotes the variable, $\mathbf{y}_{i,-j}^{obs}$ are observed variables in i, and $\epsilon_i \sim \mathcal{N}(0, \nu^2)$ is the error term. From the linear model 1, we can estimate the distribution of missing value $y_{i,j}^{mis} \sim \mathcal{N}(\mu_{i,j}, \sigma_{i,j}^2)$ as:

$$\widehat{\mu_{i,j}} = \mathbf{y}_{i,-j}^{obs} \widehat{\boldsymbol{\beta}} \tag{2}$$

$$\widehat{\sigma_{i,j}} = \widehat{\nu}. \tag{3}$$

To estimate the linear models in the presence of missing values, the MBEMMI method employs an Expectation Maximisation (EM) process, similar to Honaker and King (2010), that iteratively estimates the linear models using the sufficient statistics, Q. This process involves replacing missing values with current expectations until convergence is achieved.[3]

If linear aggregation constraints are not considered, the EM process inevitably converges to a local maximum, resulting in estimates of the missing value distributions that violate these constraints. Furthermore, the estimates are less accurate due to the neglect of valuable information embedded within the linear constraints. To incorporate the aggregations, the MBEMMI method employs singular covariance matrices within each iteration of the EM process. This approach effectively adjusts the estimates of the missing data distribution, ensuring their adherence to the linear constraints.

Specifically, analogous to the multiscale step in Holan et al. (2010), the MBEMMI method segments the hierarchical multidimensional data into Basic Constraint Units (BCU). A BCU is defined as the the smallest unit that keeps the multidimensional linear constraints. Considering the QCEW sample in Table 1 as an example, one BCU encompasses the first five rows, representing all data points within year one. Any further segmentation would disable one or more linear constraints. MBEMMI subsequently transforms these BCUs into vectors, $z_{i'}$, where i' represents the year. Conditional on the current estimates of $\widehat{\mu_{i,j}}, i \in (4i'-3, 4i'), j \in (1,2,3)$, vector $z_{i'}$ follows a multivariate normal distribution $\mathcal{N}(\mu_{i'}, \Sigma_{i'})$. The covariance matrix $\Sigma_{i'}$ exhibits singularity due to redundant information within the linear aggregations. MBEMMI then partitions the vector $z_{i'}$ into observed values $z_{i',o}$ and missing values $z_{i',m}$, leading to a subsequent mean vector $\mu_{i'} = (\mu_{i',o}, \mu_{i',m})$, and a

[3] Details of the EM process are provided in Appendix A: The Expectation Maximization Process.

covariance matrix:

$$\Sigma_{i'} = \begin{pmatrix} \Sigma_{i',oo} & \Sigma_{i',om} \\ \Sigma_{i',mo} & \Sigma_{i',mm} \end{pmatrix}.$$

In each step of the EM process, the MBEMMI method leverages the Moore-Penrose inverse $\Sigma^{+}_{i',oo}$ (Searle, 1982) to adjust the estimated distribution of missing values in each step of the EM process:[4]

$$\begin{aligned} z_{i',m} \mid z_{i',o} &\sim \mathcal{N}(\gamma_{i',m}, \Omega_{i',m}) \\ \gamma_{i',m} &= \mu_{i',m} - \Sigma_{i',mo}\Sigma^{+}_{i',oo}(z_{i',o} - \mu_{i',o}) \\ \Omega_{i',m} &= \Sigma_{i',mm} - \Sigma_{i',mo}\Sigma^{+}_{i',oo}\Sigma_{i',om}. \end{aligned}$$

With the converged distributions of missing values, MBEMMI can generate random draws that satisfy the linear constraints. To comprehensively explore the uncertainty associated with missing values, rather than drawing multiple imputations from a single converged distribution, MBEMMI adopts a Quasi-Monte Carlo bootstrapping method, similar to Honaker and King (2010). This approach creates m variations of the incomplete dataset and conducts independent EM processes to estimate m missing value distributions. MBEMMI then generates one imputation from each of these estimated distributions.[5]

This bootstrapping-initiated multiple imputation method offers several advantages:

- **Enhanced speed:** It outperforms maximum likelihood and IP methods in terms of computational efficiency (Honaker and King, 2010).
- **Parallelization:** It naturally supports "embarrassingly parallel" execution as the bootstrapped processes are independent.
- **Deterministic convergence:** The EM processes converge deterministically, eliminating the need for expert supervision.

[4] Detailed steps for incorporating multidimensional linear aggregations are outlined in Appendix B: Incorporating Multidimensional Linear Aggregation Constraints.

[5] Detailed information on the Quasi-Monte Carlo bootstrapping method is provided in Appendix C Quasi-Monte Carlo Bootstrapping Method.

As shown in Figure 1, the MBEMMI method consists of the following key steps:[6]

Step 1. Quasi-MC Bootstrapping: Bootstrap datasets $(\mathbf{Y}'_1, \mathbf{Y}'_2, \ldots, \mathbf{Y}'_m)$. For each bootstrapped dataset $\mathbf{Y}'_k$, apply step 2-6.

Step 2. Compute Q: Construct the initial sufficient statistics $Q = (\mathbf{Y}'_k)^T(\mathbf{Y}'_k)$.

Step 3. Expectation: Estimate the distribution of missing values $(\widehat{\mu}, \widehat{\Sigma})$.

Step 4. Incorporating Aggregations: Use linear aggregation constraints to correct the distribution of missing values, $(\widehat{\mu}', \widehat{\Sigma}')$.

Step 5. Maximization: Construct new sufficient statistics Q'. If Q' converged, continue to step 6, otherwise repeat step 3-5.

Step 6. Imputation: Obtain converged distribution $(\widehat{\mu}^*, \widehat{\Sigma}^*)$, draw one imputation. Insert imputation in original dataset $\mathbf{Y}$, obtain imputed dataset $\mathbb{Y}_k$.

4 Parallel Sequential Imputation Scheme

Due to the independence and automatic convergence of its multiple imputation processes, the MBEMMI method can be efficiently scaled for large hierarchical multidimensional data through a Parallel Sequential Imputation (PSI) blocking scheme.

Consider the 2012-2016 Florida QCEW data as an example (Table 2). Imputing 2,157 industries while simultaneously accounting for linear aggregations is not only computationally demanding but also poses potential challenges related to inverting large covariance matrices. Fortunately, the QCEW data utilizes a tree-like NAICS code structure with five hierarchical levels, enabling its segmentation into smaller, more manageable blocks. Each block encompasses a single coarser-resolution industry along with its immediate sub-industries, similar to the sample in Table 1.

Since highly correlated industries are already grouped under the same coarser-resolution industry within the NAICS code structure, and research has demonstrated that linear

[6] Also see Algorithm 1 in Appendix D: The MBEMMI and PSI Algorithms.

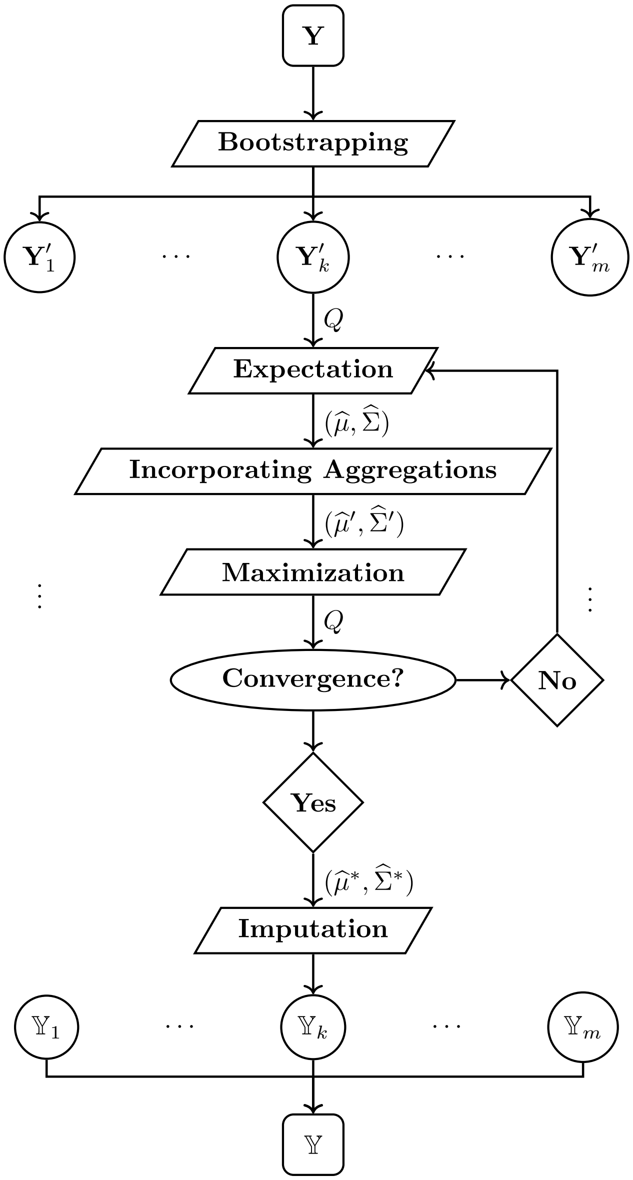

Figure 1: The MBEMMI Method

aggregations capture most inter-block correlations (Holan et al., 2010), we can treat these smaller blocks as independent entities. This permits their separate imputation, significantly enhancing data manageability.

Furthermore, Table 2 reveals that higher (coarser resolution) levels tend to exhibit fewer incomplete industries. This is because as the level increases, vulnerable industries become increasingly encompassed within aggregations, necessitating less suppression. This pattern proves advantageous for imputation purposes, as information about missing values remains embedded within higher-level aggregations, albeit in a mixed form. By imputing blocks in a top-down, level-by-level manner, we effectively transfer information about missing values to lower levels where it is most needed. This scheme introduces additional information into the imputation process, leading to improved results.

The PSI scheme involves sequential imputation of QCEW data level by level, with parallel imputation of separate blocks within each level. Upon completion, we generate a single imputation for the entire dataset and can repeat the scheme in parallel m times to obtain m imputations.[7]

MBEMMI uniquely stands as the sole multiple imputation method that fully supports the PSI scheme. The BMMI method necessitates automatic tools for convergence detection and the employment of certain prior rules to guarantee the quality of imputations derived from parallel processes rather than a single Markov chain. On the other hand, the EMB method cannot account for linear aggregations, weakening the assumption of inter-block independence and negatively impacting imputation accuracy.

The subsequent section will compare the performance of these multiple imputation methods using real QCEW data.

5 Validating the Algorithm

This section validates the MBEMMI method in terms of accuracy and speed, drawing comparisons with two prominent alternatives, BMMI and EMB, using real QCEW data. We start with a detailed examination of the QCEW sample presented in Table 1, followed by

[7] Detailed steps of the PSI scheme are outlined in Algorithm 2 in Appendix D: The MBEMMI and PSI Algorithms.

an exploration of larger QCEW datasets using the PSI scheme. Comprehensive performance statistics will be reported for each dataset.

5.1 Validation using a small QCEW sample

Table 3 presents the QCEW sample after imputation using the MBEMMI method. Each missing quarterly employment count was imputed $m = 10$ times, resulting in the generation of ten completed datasets. The missing annual totals for Series 1 and 2 in year 2 can be calculated by summing the respective quarterly counts within that year. Notably, all linear aggregation constraints hold within these completed datasets.

Table 3: MBEMMI imputed QCEW sample in Table 1. Each missing cell was imputed $m = 10$ times. Only 95% confidence intervals are shown.

	Series 1	Series 2	Series 3	Total
year1.q1	20	414	484	918
year1.q2	24	412	493	929
year1.q3	25	404	508	937
year1.q4	23	415	527	965
year1.a	92	1,645	2,012	3,749
year2.q1	9	262	540	811
year2.q2	**(6, 24)**	**(258, 276)**	557	839
year2.q3	**(11, 28)**	**(293, 310)**	510	831
year2.q4	**(6, 26)**	**(314, 334)**	528	868
year2.a	**-**	**-**	2,135	3,349
year3.q1	**(6, 17)**	**(507, 519)**	676	1,200
year3.q2	21	495	684	1,200
year3.q3	20	468	665	1,152
year3.q4	**(21, 32)**	**(482, 493)**	703	1,217
year3.a	79	1,964	2,728	4,769
year4.q1	32	476	645	1,153
year4.q2	30	473	652	1,155
year4.q3	31	484	686	1,200
year4.q4	30	553	723	1,306
year4.a	123	1,986	2,706	4,814
year5.q1	36	538	630	1,205
year5.q2	41	502	661	1,204
year5.q3	45	500	657	1,202
year5.q4	48	514	639	1,200
year5.a	170	2,054	2,587	4,811

Due to confidentiality restrictions surrounding suppressed missing values in the QCEW sample, direct display of imputed values or comparison with true values is not feasible. However, 95% confidence intervals, presented in bold numbers within missing cells, offer valuable insights. Notice the general alignment of these intervals with observed values immediately preceding or succeeding missing values. This suggests that the imputations successfully preserve the employment time series' trends and that imputed values do not exhibit substantial deviations from observed values. The width of confidence intervals varies according to uncertainty introduced by missing values. For example, intervals in Year 2 tend to be wider than those in Year 3 due to a greater prevalence of missing values in Year 2, including two missing annual totals. Increased missing values equate to greater information loss, leading to greater uncertainty during imputation.

Directly evaluating the accuracy of our method poses a challenge due to the inherent confidentiality of suppressed data, preventing access to true values. To circumvent this obstacle and rigorously assess the method's effectiveness, we've employed a strategic approach: replacing all values in Table 3 with data from alternative industries unaffected by suppression. By applying the same suppression patterns from Table 3 to this new data, we create a scenario where the "true" suppressed values are no longer confidential, enabling explicit validation of the method.

Imputations for the new QCEW sample were generated using MBEMMI, BMMI, and EMB, each method producing $m = 10$ imputations. Figure 2 displays the complete Series 1 and 2, accompanied by 95% confidence intervals of imputed values depicted as error bars. Visually, error bars in the top figure appear smaller than those in the bottom figure, although they are of similar magnitudes. This phenomenon stems from the constraints imposed by industry-wise aggregations. When an imputation in Series 1 at time i, denoted as $\tilde{y}_{i,1}$, experiences an increase of s, the corresponding imputation $\tilde{y}_{i,2}$ must decrease by s to maintain the linear aggregation. This inherent property results in proportionally wider error bars within smaller industries (like Industry 2) and narrower error bars within larger industries (like Industry 1).

In Series 1, both MBEMMI and BMMI demonstrate high accuracy and capture the true values within their relatively narrow error bars. Conversely, EMB exhibits larger

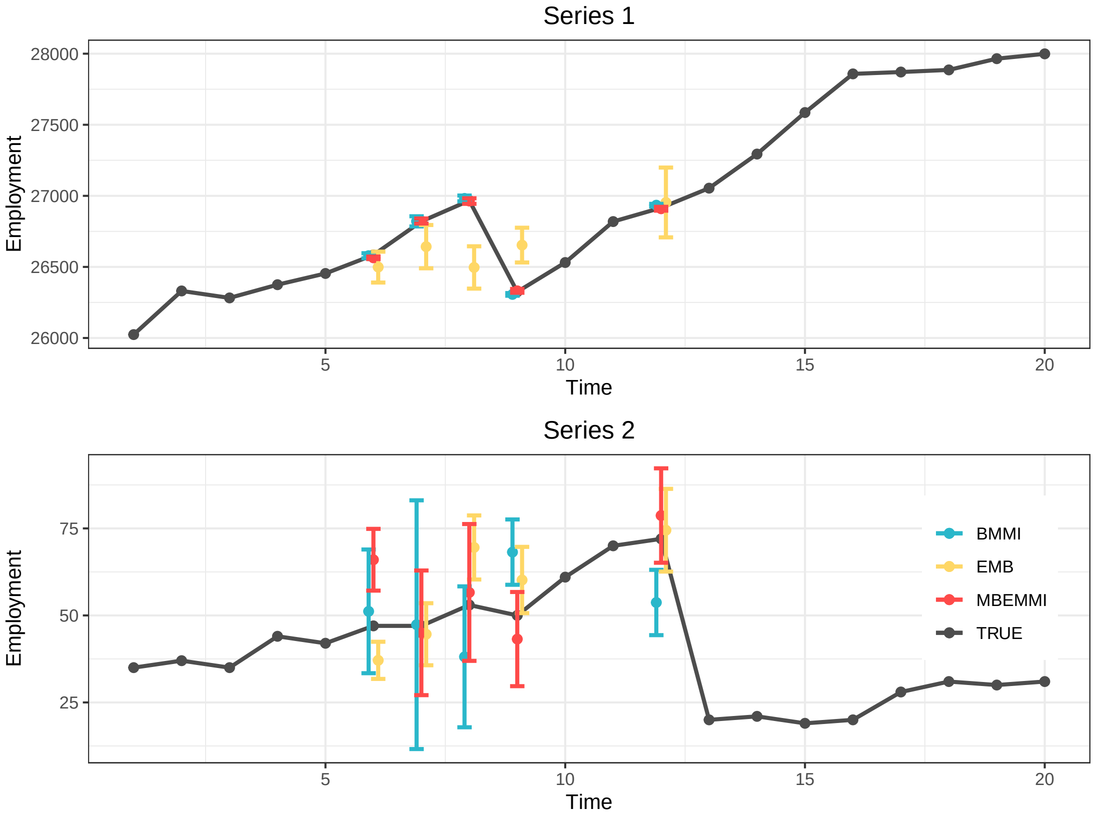

Figure 2: Imputed QCEW sample in Table 1. For confidentiality reasons, the values in the sample are replaced with disclosed data while the suppressions are unchanged.

error bars, and three of its imputations deviate from the true values. In Series 2, the accuracy of all three methods aligns more closely, with MBEMMI maintaining a slight edge. MBEMMI's error bars miss a single true value, while BMMI misses two, and EMB misses three. The discrepancy in EMB's performance between Series 1 and 2 can be attributed to its underlying assumption that missing values adhere to existing trends. This assumption holds true for Series 2 but proves less accurate for Series 1's where there is a sharp level shift in the midst of the cluster of missing values. Consequently, EMB performs better in scenarios characterized by limited variability within series. Conversely, MBEMMI and BMMI demonstrate good performance irrespective of trends, as they leverage strength from the linear aggregation constraints.

5.2 Validation using the full Florida QCEW

We now validate the methods using the full QCEW data for Florida. As discussed in section 4, we use PSI scheme to separate the large hierarchical multidimensional dataset into small blocks to avoid imputing the whole dataset at once. The individual blocks are imputed in a parallel-sequential manner that increases both the imputation speed and the resistance to failures caused by large datasets.

Table 4 summarizes the blocks resulting from PSI. Out of 1,114 blocks, 69 have missing values and require imputation. However, most of these blocks contain time series that have more than 60% of their values missing. In fact, 28 of the blocks contain completely missing series and 8 contain series missing more than 80% values. As the severely missing series contain little information for meaningful imputations (Rubin, 1996), we focus here on the 28 blocks that have missing rates of less than 60% of the values.

We impute these 28 blocks using MBEMMI, BMMI, and EMB, with $m = 10$ imputations for each method and these imputations are compared to the true values. We define the metric "τ% hit-rate" to measure the accuracy of the MI methods as the percentage of imputed values that are within τ% of the true values:

$$\psi_\tau^p = \frac{\sum_{i,j \in \mathbb{M},k} 1_{\frac{|\tilde{y}_{i,j}^{p,k} - y_{i,j}|}{y_{i,j}} \leq \tau\%}}{N_{\mathbb{M}} \times m}$$

Table 4: Blocks determined by PSI from the 2012-2016 Florida QCEW data. The three columns are (1) all possible blocks, (2) blocks with suppressed values, (3) blocks that do not contain severely suppressed series, i.e. series have more than 60% values suppressed. A block in the k-digit NAICS code level contains one k-digit industry and all of its (k+1)-digit sub-industries.

Level	Blocks	Incomplete	$(0\%, 60\%]$
2-digit	25	1	0
3-digit	94	6	3
4-digit	316	21	9
5-digit	679	41	16
Total	1,114	69	28

where $p \in \{MBEMMI, BMMI, EMB\}$ denotes the multiple imputation method, $k \in (1, 2, ..., m)$ is the imputation indicator, and $\mathbb{M}$ is the collection of missing $y_{i,j}$ values.

The hit-rates ψ^p_τ, for $\tau = \{1\%, 2\%, 5\%, 10\%\}$ for each method on the 28 QCEW samples are shown in Table 5. In every case, the MBEMMI method hits more targets than either BMMI or EMB. Strikingly, 10.58% of the MBEMMI imputations are within the 1% interval of the true values, which is more than double the hit rate of the BMMI method and more than triple the hit rate for the EMB method. The BMMI method has higher hit-rates than the EMB method as it can account for the linear aggregation constraints while the EMB method cannot.

Table 5: Percentage of imputed values within 1%, 2%, 5%, 10% of the true values. The QCEW samples are separated from the Florida QCEW dataset using the PSI scheme discussed in Section 4. They include all individual blocks that consist of more than one sub-industries while the suppression rates of any sub-industries do not exceed 60%.

	QCEW Samples			
Method	<1%	<2%	<5%	<10%
MBEMMI	**10.58%**	**15.92%**	**25.38%**	**37.31%**
BMMI	4.19%	8.04%	18.19%	29.69%
EMB	2.62%	5.62%	11.92%	20.19%

Table 6 shows the average speed of each imputation method. The EMB method uses on average 0.02 seconds to produce one imputed QCEW sample, the MBEMMI method uses around 0.63 seconds, and the BMMI method uses 5.39 seconds on average due to the 8,000 period burn-in which is necessary for the Markov chains to converge. The 95% confidence interval based on the 10 imputations is shown in parentheses.

Table 6: Average speed and 95% confidence intervals per one QCEW sample imputation. The burn-in period for the BMMI method is set to 8,000 to ensure convergence in QCEW samples such as in Table 1. All tests run with R version 4.2.0 on an Apple MacBook M1 Pro, with 8 cores, and 16 GB of memory.

One QCEW Sample Imputation	
Method	Avg. Time (95% CI)
MBEMMI	0.63sec (0.41sec, 0.85sec)
BMMI	5.39sec (4.89sec, 5.89sec)
EMB	**0.02sec (0.01sec, 0.02sec)**

5.3 Randomly suppressed QCEW datasets

To further explore the validity of the MBEMMI method in imputing large hierarchical multidimensional datasets, we construct 10 randomly suppressed QCEW datasets. For each dataset, we randomly suppress the fully-observed confidential Florida QCEW dataset, then conduct recursive secondary suppression (Cohen and Li, 2006) to protect the initial suppressions from being computed from the linear aggregations. The random suppression datasets do not have severely missing series so we do not encounter any problematic series that may break the NAICS structure. Thus, we can focus on how the MI methods work on the entire large datasets.[8]

Following the PSI scheme, each randomly suppressed dataset is imputed $m = 10$ times in a parallel-sequential manner. The pooled hit-rates of each MI method are shown in Table 7. We note that all methods have higher hit-rates than in the QCEW samples because the missing rates are lower in the random suppression datasets. The MBEMMI method has the highest hit-rates in all categories and performs especially well in the high accuracy categories ψ_1 and ψ_2 where it has hit-rates about twice as high as the other methods. The EMB method performs similar to the BMMI method because the suppressed values were randomly selected and are less likely to be small values that deviate from the existing trend, in which case, the EMB method works better.

[8] See Table 9 in Appendix E: Statistics of the Random Suppression Datasets for a summary of the results from the random suppression simulations.

Table 7: Percentage of imputed values within 1%, 2%, 5%, 10% of the true values. The ten randomly suppressed Florida QCEW datasets are obtained by applying random primary suppression and recursive secondary suppression on the true (unsuppressed) Florida QCEW data.

Random Suppression				
Method	<1%	<2%	<5%	<10%
MBEMMI	**15.52%**	**23.12%**	**38.68%**	**53.96%**
BMMI	8.31%	14.3%	27.31%	40.87%
EMB	7.56%	14.51%	31.46%	48.87%

We also report the average speeds of the MI methods in Table 8. The EMB method uses around 0.01 seconds to impute a single block once, while the MBEMMI method takes 0.29 seconds. Both methods are much faster than the BMMI method which takes over 6 seconds. The random suppression datasets also allow us to test the speeds of the MI methods on full datasets. To create ten imputed QCEW datasets, the EMB method takes 1.85 minutes and the MBEMMI method uses on average 4.79 minutes. The BMMI method requires on average 51.23 minutes to compute full dataset imputations.

Table 8: Average speed and 95% confidence intervals per one random suppression block imputation and per ten random suppression dataset imputations. The burn-in period for BMMI method is set to 8,000 as it is necessary to ensure convergence in QCEW samples such as in Table 1. The PSI scheme is applied to the random suppression datasets. All tests are in R version 4.2.0 on an Apple MacBook M1 Pro, with 8 cores, and 16 GB memory.

One Block Imputation	
Method	Avg. Time (95% CI)
MBEMMI	0.29sec (0.27sec, 0.31sec)
BMMI	6.44sec (6.35sec, 6.54sec)
EMB	**0.01sec (0.01sec, 0.01sec)**
Ten Full Data Imputations	
Method	Avg. Time (95% CI)
MBEMMI	4.79min (3.96min, 5.61min)
BMMI	51.23min (50.16min, 52.29min)
EMB	**1.85min (0.37min, 4.95min)**

These tests on samples and the full size QCEW datasets demonstrate that the new MBEMMI method can impute missing values in large hierarchical multidimensional datasets accurately and fast.

6 Empirical Application: Average Weekly Wage

In this section we apply the MBEMMI method to a panel data model of average weekly wages by industry as a function of employment level and industry establishment counts from the QCEW data along with the macroeconomic variables GDP growth rates, inflation rates, and the unemployment rates. The model is specified as:

$$\begin{aligned} Wage_{i',j} =& \beta_1 Employment_{i',j} + \beta_2 Establishment_{i',j} + \beta_3 GDP_{i'} + \\ & \beta_4 Inflation_{i'} + \beta_5 Unemployment_{i'} + \upsilon_{i'} + \varphi_j + \epsilon_{i',j} \end{aligned} \tag{4}$$

where i' denotes the year, j represents the industry, $\upsilon_{i'}$ and φ_j are fixed time and unit effects, and $\epsilon_{i',j}$ is the idiosyncratic error. We assume that all classical assumptions for the panel data model hold.

We estimate the model over the QCEW sample from Table 1 that includes data from 2012 through 2016 for three industries. Two of those industries have some suppressed employment values and the third has complete data. Although the data released by the BLS does have suppressions, we also have the true unsuppressed data values so that we can test the efficiency of the MBEMMI method.

We use MBEMMI to generate $m = 10$ imputed QCEW samples and then estimate the fixed effect model 4 on every imputed dataset and obtain 10 sets of estimates. Then we pool the results using Rubin's MI pooling rules (Rubin, 1987)[9] to compute the imputation adjusted point estimates and standard errors for the estimated parameters of the model. We also estimate the model using the unsuppressed data as well as the list-wise deleted data.

Figure 3 shows the point estimates and their 95% confidence intervals for all five coefficients of the model. The black intervals represent the estimation using the true unsuppressed data, the blue intervals represent estimations where records with suppressed values are deleted (list-wise deletion or complete case estimation), and the red intervals represent

[9] Following Rubin's rules (Rubin, 1987), the pooled point estimate is obtained through $\tilde{\beta}_\lambda = \frac{1}{m}\sum_{k=1}^{m}\widehat{\beta}_\lambda^k$, where λ indicates covariates. The pooled variance (T_λ) is a combination of the average within-imputation variance ($\bar{U}_\lambda$) and the between-imputation variance (B_λ): $T_\lambda = \bar{U}_\lambda + (1+\frac{1}{m})B_\lambda$, where the average within-imputation variance is $\bar{U}_\lambda = \frac{1}{m}\sum_{k=1}^{m} U_\lambda^k$, and the between-imputation variance is $B_\lambda = \frac{1}{m-1}\sum_{k=1}^{m}(\widehat{\beta}_\lambda^k - \bar{\beta}_\lambda)^2$.

the estimates from the MBEMMI imputed cases and the Rubin adjustments. The intervals are fairly wide in this illustration because we are using a relatively small sample for a panel data model.

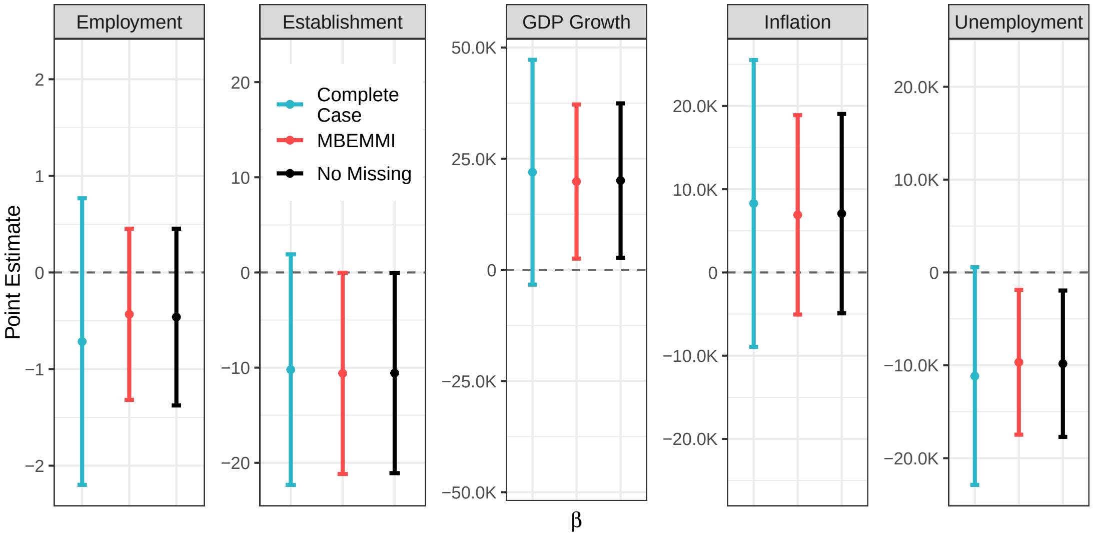

Figure 3: Point estimate and 95% confidence intervals. The black intervals are from unsuppressed data with no missing values, blue intervals are complete case estimation of the suppressed data, red intervals are from the MBEMMI imputed data.

We can learn from the Figure that none of the complete case estimates (blue) are significantly different from zero and the bias in these estimates appears to be great compared to the true data estimation (black). The MBEMMI results (red) are nearly identical to the true estimates (black) and they appear to be unbiased and to correctly recover the standard errors from the true model. Consequently, unlike the complete cases estimates, the MBEMMI results successfully rejects the null hypothesis $\beta = 0$ for variables *Establishment*, *GDP*, and *Unemployment* at 95% confidence.

This small application illustrates that MBEMMI can assist researchers in obtaining estimates that closely resemble those derived from the true data estimates with no missing values, and subsequently results in more accurate and reliable statistical inferences when missing data present.

7 Concluding Remarks

In this study, we address the difficulties in imputing data in large hierarchical multidimensional datasets with linear aggregation constraints. Although such datasets are becoming increasingly available in economcis, they pose significant challenges for existing multiple imputation methods which either fail to take linear aggregations into account, or are not fast enough for practical implementation in empirical research.

In this paper we introduce the Multidimensional Bootstrapping Expectation Maximization Multiple Imputation (MBEMMI) method, which employs singular mulitvariate normal distributions to account for the multidimensional linear constraint structure and uses an EM algorithm along with a Parallel-Sequential Imputation (PSI) scheme to facilitate rapid imputation of large datasets. The method is fast and capable of imputing large datasets while achieving comparable or superior accuracy compared to existing methods.

Using real-world datasets and an empirical application, we demonstrate that the MBEMMI method outperforms the leading alternatives in both accuracy and speed. Measured in terms of how closely imputed values match true but undisclosed data, MBEMMI is about twice as accurate in imputing the Florida QCEW samples and the randomly suppressed datasets than alternative methods. In addition, MBEMMI is approximately ten times faster than the alternative multiple imputation methods able to take the linear aggregations into account. An application of MBEMMI on a panel data estimation model with suppressed data shows that the MBEMMI method helps researchers obtain unbiased estimates and make correct inferences.

Future research will explore relaxed distributional assumptions and increased flexibility in the types of constraints that can be handled. We plan to develop a versatile R package to broaden MBEMMI's usability and accessibility for researchers dealing with large and complex datasets.

References

Aidara, C. A. T. (2013). Bootstrap variance estimation for complex survey data: a quasi monte carlo approach. *Sankhya B*, 75(1):29–41.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

Batista, G. E., Monard, M. C., et al. (2002). A study of k-nearest neighbour as an imputation method. *His*, 87(251-260):48.

Beaton, A. E. (1964). The use of special matrix operators in statistical calculus. *ETS Research Report Series*, 1964(2).

BLS (2017). The qcew hand book of methods. *U.S. Bureau of Labor Statistics. Office of Publications and Special Studies*.

Cohen, S. and Li, B. T. (2006). A comparison of data utility between publishing cell estimates as fixed intervals or estimates based upon a noise model versus traditional cell suppression on tabular employment data december 2006.

Cranmer, S. J. and Gill, J. (2013). We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, 43(2):425–449.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, New York, NY, USA.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Holan, S. H., Toth, D., Ferreira, M. A., and Karr, A. F. (2010). Bayesian multiscale multiple imputation with implications for data confidentiality. *Journal of the American Statistical Association*, 105(490):564–577.

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014). Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32(3):375–386.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1):49–69.

Kolenikov, S. (2007). Applications of quasi-monte carlo methods in inference for complex survey data. *Proceedings of the Survey Reseacrh Methods Section of ASA*.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, pages 287–296.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Matoušek, J. (1998). On the l2-discrepancy for anchored boxes. *Journal of Complexity*, 14(4):527–556.

Muirhead, R. J. (1982). *Aspects of multivariate statistical analysis.* John Wiley & Sons, Inc., New York, NY, USA.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R.* CRC press.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, pages 87–94.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley, New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, 33(4):545–571.

Searle, S. R. (1982). *Matrix algebra useful for statistics (wiley series in probability and statistics)*. Wiley-Interscience.

Siotani, T., Fujikoshi, Y., and Hayakawa, T. (1985). *Modern multivariate statistical analysis, a graduate course and handbook*. American Sciences Press, Columbus, Ohio, USA.

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4):784–802.

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Teytaud, O., Gelly, S., Lallich, S., and Prudhomme, E. (2006). Quasi-random resamplings, with applications to rule extraction, cross-validation and (su-) bagging. In *Dans International Workshop on Intelligent Information Access III A*.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Supplementary Material

A The Expectation Maximization Process

Expectation Step

In the expectation step, MBEMMI fills-in missing cells of the original dataset Y with their conditional expectations, based on the current estimates of the sufficient statistics of bootstrapped data.

The sufficient statistics are $Q = (Y_k')^T(Y_k')$, where Y_k' is the kth bootstrapped dataset whose aggregation values are excluded to avoid perfect multi-collinearity, and the first column of Y_k' is a vector of ones. To illustrate, suppose the bootstrapped dataset Y_k' is:

$$Y_k' = \begin{pmatrix} 1 & y_{4,1} & S & \cdots & S & A_4 \\ 1 & S & y_{7,2} & \cdots & S & A_7 \\ 1 & S & S & \cdots & y_{3,N_j} & A_3 \\ 1 & y_{2,1} & y_{2,2} & \cdots & y_{2,N_j} & A_2 \\ 1 & S & y_{7,2} & \cdots & S & A_7 \\ 1 & S & S & \cdots & y_{3,N_j} & A_3 \\ 1 & y_{5,1} & y_{5,2} & \cdots & y_{5,N_j} & A_5 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \end{pmatrix}$$

where, for instance, the first subscript in $y_{4,1}$ denotes that the first row of bootstrapped dataset Y_k' was from quarter $i = 4$ from the original dataset Y'. A_4 denotes the auxiliary variables which are always observed. The sufficient statistics for data Y_k' are computed as

$Q_k = (Y_k')^T(Y_k')$. It is convenient to rewrite the sufficient statistics as

$$Q_k^* = \begin{pmatrix} -1 & \hat{\mu}_1 & \hat{\mu}_2 & \cdots & \hat{\mu}_{N_j} & \hat{\mu}_A \\ \hat{\mu}_1 & \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1N_j} & \hat{\sigma}_{1A} \\ \hat{\mu}_2 & \hat{\sigma}_{12} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2N_j} & \hat{\sigma}_{2A} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\mu}_{N_j} & \hat{\sigma}_{1N_j} & \hat{\sigma}_{2N_j} & \cdots & \hat{\sigma}_{N_j}^2 & \hat{\sigma}_{N_jA} \\ \hat{\mu}_A & \hat{\sigma}_{1A} & \hat{\sigma}_{2A} & \cdots & \hat{\sigma}_{N_jA} & \hat{\sigma}_A^2 \end{pmatrix},$$

where, in the first EM iteration, missing values in Y_k' are replaced with column means.

The MBEMMI algorithm executes estimations by using a SWEEP operator $\theta(s)$ (Beaton, 1964). Given an $1 \times (1 + N_j + P)$ input vector s that consists of only 1's and 0's, where $(1 + N_j + P)$ is the number of columns in Q_k^* and P is the number of auxiliary variables, a SWEEP operator $\theta(s)$ will operate on the elements of Q_k^* and obtain corresponding estimates. The intuition of the SWEEP operator is to transform the Q_k^* matrix and obtain parameter estimates of all of the functions whose dependent variable is marked as 1 in s and explanatory variables are 0 in s.

Consider a simple example that has only two variables so that Q^* is

$$Q^* = \begin{pmatrix} -1 & \hat{\mu}_1 & \hat{\mu}_2 \\ \hat{\mu}_1 & \hat{\sigma}_1^2 & \hat{\sigma}_{12} \\ \hat{\mu}_2 & \hat{\sigma}_{12} & \hat{\sigma}_2^2 \end{pmatrix}.$$

The SWEEP operator with $s = (0, 1, 0)$ will transform Q^* to $\theta(s = \{0, 1, 0\})$

$$\theta(s = \{0,1,0\}) = \begin{pmatrix} -1 - \frac{(\hat{\mu}_1)^2}{\hat{\sigma}_1^2} & \frac{\hat{\mu}_1}{\hat{\sigma}_1^2} & \hat{\beta}_0 = \hat{\mu}_2 - \frac{\hat{\sigma}_{12}\hat{\mu}_1}{\hat{\sigma}_1^2} \\ \frac{\hat{\mu}_1}{\hat{\sigma}_1^2} & -\frac{1}{\hat{\sigma}_1^2} & \hat{\beta}_1 = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1^2} \\ \hat{\mu}_2 - \frac{\hat{\sigma}_{12}\hat{\mu}_1}{\hat{\sigma}_1^2} & \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1^2} & \hat{\sigma}_{2|1}^2 = \hat{\sigma}_2^2 - \frac{(\hat{\sigma}_{12})^2}{\hat{\sigma}_1^2} \end{pmatrix}.$$

Note that $s = \{0, 1, 0\}$ means we are estimating $y_{i,2} = \beta_0 + \beta_1 y_{i,1} + \epsilon$. The values obtained in the third column above are the coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ and the variance $\hat{\sigma}_{2|1}^2$.

Using the SWEEP operators, the expectations of missing values are estimated by equation

$$E(y_{i,j}) = \mathbf{y}_i^{obs}\theta\{1 - M_i\}_j^t,$$

where $s = (1 - M_i)$ is the observed value indicator of quarter i, $\mathbf{y}_i^{obs}$ is the $1 \times (1 + N_j + P)$ vector of observations at time i with zeros in the missing cells, and $(1 + N_j + P) \times 1$ vector $\theta\{1 - M_i\}_j^t$ denotes that we are using the j^{th} column of swept matrix θ obtained in the t^{th} iteration. The transformation $Q \rightarrow Q^*$ can be seen as a variation of a SWEEP operator applied on $s = \{1, 0, 0\}$. Thus, there is no need to SWEEP on the first row and column again, allowing the first element of s to always be 0. Continuing with the previous simple example, this equation is computing $E(y_{i,j}) = \beta_0 + \beta_{-j} \times y_{i,-j}$, where $-j$ stands for other numbers in $\{1, 2, 3\}$ but not j. The variance of the missing values are obtained as

$$\widehat{\sigma}^2_{2|1} = \theta\{1 - M_i\}_{i,j}^t$$

where the subscript of $\theta\{1 - M_i\}_{i,j}^t$ denotes the value of the i^{th} row and j^{th} column of swept matrix θ.

To utilize the aggregation constraints from the original dataset Y, instead of filling in the expectations in the missing cells of bootstrapped dataset Y_k', we first compute the sufficient statistics Q_k from bootstrapped dataset Y_k' and then use the sweep operator θ according to the locations of the suppressed values in the original dataset Y.

For each observation $\mathbf{y}_i$ in original dataset Y', the filled-in row vector $\widehat{\mathbf{y}}_i^E$ and covariance matrix of missing values $\Sigma^E_{i|\mathbf{y}_i^{obs}}$ will be

$$\begin{aligned}\widehat{\mathbf{y}}_i^E &= \mathbf{y}_i^{obs} + M_i \cdot (\mathbf{y}_i^{obs}\theta\{1 - M_i\}^t)\\ \Sigma^E_{i|\mathbf{y}_i^{obs}} &= M_i'M_i \cdot \theta\{1 - M_i\}^t\end{aligned}$$

where $\theta\{1 - M_i\}^t$ is the whole swept matrix θ obtained in the t^{th} iteration and the "$\cdot$" denotes the inner product operator.

Maximization Step

In the maximization step, we reconstruct the sufficient statistics as:

$$Q' = \sum_i [(\widehat{\mathbf{y}}_i^E)^T(\widehat{\mathbf{y}}_i^E) + \Sigma^E_{i|\mathbf{y}_i^{obs}}],$$

The sufficient statistics Q' will be used in the next E-Step to generate new expectations and variances for missing values of the original dataset Y'. The EM process continues until the sufficient statistics Q converges.

B Incorporating Multidimensional Linear Aggregation Constraints

After the E-step the expectations of suppressed values have been estimated using $Q_A(t)$ from bootstrapped dataset $Y'_A(t)$, which was previously stripped of the aggregation constraints in the raw data file Y, and substituted back into the appropriate locations in Y'. The MBEMMI method then enforces the binding constraints to produce the updated dataset, say, $\tilde{Y}'(t)$. We then use the stored mapping vector to map the suppressed values back into an updated bootstrapped dataset, $Y'_A(t{+}1)$. Since the observation-by-observation bootstrapping will not generally select all four quarters of the same year into the sampled data $Y'_A(i+1)$, the aggregation constraints will not bind in the bootstrapped data. In the this step we reimpose the multidimensional linear aggregation constraints using a technique follows the multiscale step of Holan et al. (2010) and construct a new sufficient statistics matrix that incorporates these constraints.

The step starts by transforming quarterly data $y_{i,j}$, quarterly aggregations q_i and annual aggregations $a_{i',j}$ into a yearly vector

$$\begin{aligned} z_{i'} = (&y_{4i'-3,1}, ..., y_{4i',1}, y_{4i'-3,2}, ..., y_{4i',1}, ..., y_{4i'-3,N_j}, ..., y_{4i',N_j}, \\ &q_{4i'-3}, ..., q_{4i'}, a_{i',1}, ..., a_{i',N_j})' \end{aligned}$$

where $i' = \{1, 2, 3, ..., \frac{N_i}{4}\}$ represents the index for years.

Using year 1 as an example, the transformation places the values of the first five rows of dataset Y into a column vector z_1. z_1 uses only observable aggregation values.

Define the operator matrix H:

$$H = \begin{pmatrix} & I_{4N_j} & & \\ I_4 & I_4 & \cdots & I_4 \\ & I_{N_j} \otimes 1'_4 & & \end{pmatrix}$$

where I_n is the $n \times n$ identity matrix, 1_n is a $n \times 1$ vector of ones and $\otimes$ denotes the Kronecker product. The dots in this operator matrix represent that there are N_j identity matrices of size 4×4. For z_1, since the aggregation values are not complete, the corresponding operator matrix is

$$H = \begin{pmatrix} & I_{4N_j} & & \\ I_4 & I_4 & \cdots & I_4 \\ & I_{(N_j-2)\times N_j} \otimes 1_4' & & \end{pmatrix}$$

where $I_{(N_j-2)\times N_j}$ is a (N_j-2)-by-N_j matrix transformed from identity matrix I_{N_j} by deleting the corresponding row of I_{N_j} whenever the annual total is not observed. In this example, these are rows 4 and 5.

Given the appropriate operator matrix H, we obtain the mean vector $\mu_{i'}$ and covariance matrix Σ of $z_{i'}$ as

$$\begin{aligned} \mu_{i'} &= H\theta_{i'} \\ \Sigma &= HVH', \end{aligned}$$

where

$$\begin{aligned} \theta_{i'} &= (E(y_{4i'-3,1}), ..., E(y_{4i',1}), E(y_{4i'-3,2}), ..., E(y_{4i',2}), ..., E(y_{4i'-3,N_j}), ..., E(y_{4i',N_j})), \\ V &= diag(\hat{\sigma}_1^{2^E}, \hat{\sigma}_1^{2^E}, \hat{\sigma}_1^{2^E}, \hat{\sigma}_1^{2^E}, \hat{\sigma}_2^{2^E}, \hat{\sigma}_2^{2^E}, \hat{\sigma}_2^{2^E}, \hat{\sigma}_2^{2^E}, ..., \hat{\sigma}_{N_j}^{2^E}, \hat{\sigma}_{N_j}^{2^E}, \hat{\sigma}_{N_j}^{2^E}, \hat{\sigma}_{N_j}^{2^E}). \end{aligned}$$

In $\theta_{i'}$, the expectations are actual values if the corresponding values are not suppressed, and they are expectations obtained from the E-step if the corresponding values are suppressed. It is possible to expand V to include covariances $\hat{\sigma}^E$ in the off-diagonal positions but our computations indicate that there is little gain from this so we use a diagonal V for computational convenience.

To investigate the posterior distribution of the missing values conditional on observed values, we partition Σ into 4 blocks:

$$\Sigma = \begin{pmatrix} \Sigma_{oo} & \Sigma_{om} \\ \Sigma_{mo} & \Sigma_{mm} \end{pmatrix},$$

where Σ_{oo} is the covariance matrix of observed values, Σ_{mm} is the covariance matrix for missing values and $\Sigma_{om} = \Sigma_{mo}$ is the covariance matrix of the missing values with the observed values. Due to the aggregation constraints, Σ_{oo} is usually singular, in which cases, we use the Moore-Penrose inverse

$$\Sigma_{oo}^{+} = P^{*}(D^{*})^{-1}P^{*'}, \tag{5}$$

where D^{*} is a diagonal matrix with the non-zero eigenvalues of Σ_{oo} on its diagonal and P^{*} are the corresponding eigenvectors.

Using standard properties of normal distributions with singular covariance matrices (Muirhead, 1982; Siotani et al., 1985), the posterior distribution of missing values $z_{i',m}$ conditional on observed values $z_{i',o}$ is

$$z_{i',m} \mid z_{i',o} \sim \mathcal{N}(\gamma_{i',m}, \Omega_{m}),$$

where the mean vector $\gamma_{i',m}$ and covariance matrix Ω_{m} are given by

$$\begin{aligned}\gamma_{i',m} &= \mu_{i',m} + \Sigma_{mo}\Sigma_{oo}^{+}(z_{i',o} - \mu_{i',o}),\\ \Omega_{m} &= \Sigma_{mm} - \Sigma_{mo}\Sigma_{oo}^{+}\Sigma_{om}.\end{aligned}$$

At the completion of the step the posterior distributions of the missing variables based upon a particular bootstrapped dataset and incorporating the aggregation constraints is now available.

C Quasi-Monte Carlo Bootstrapping Method

We use an observation-by-observation quasi-Monte Carlo bootstrapping technique in MBE-MMI to generate unique independent multiple imputations. The Bootstrapping-based Expectation Maximization method (EMB) developed by Honaker and King (2010) is a generalization of the bootstrapping technique for missing data problems and is preferred over the more complicated process of drawing μ and Σ from their posterior density used in Imputation-Posterior (IP) methods. Given standard regularity conditions and as the sample size grows larger, bootstrapped data will have approximately the same properties as the original data (Efron and Tibshirani, 1994) and has lower order asymptotics than the parametric approaches used in IP and EM with importance re-sampling (EMis) (Honaker and King, 2010). These advantages allow us to use a bootstrapping technique to obtain similar random draws from the posterior in a relatively shorter time.

Starting from the input data file represented in Table 1, we first eliminate the annual total rows ($a_{i',j}$'s) and the quarterly totals (q_i's) to get a $N_i \times N_j$ array of $y_{i,j}$'s. Next, to each row we add a vector of auxiliary variables, A_i, that will be used in the expectation step to improve estimates of missing data. These auxiliary variables include the number of establishments in each industry and basis functions of time created via polynomials, LOESS, splines or wavelets. Denote this modified input dataset as Y'. The aggregated values are excluded from Y' to avoid perfect multi-colinearity in the OLS estimation of the E-step. However, these aggregate values are essential to the incorporating aggregation step so they are kept aside in original dataset Y along with mapping indicators to allow us map the aggregations in Y back to the detailed values in Y'.

The bootstrapping process randomly picks one quarterly observation at a time with replacement from Y' and stacks them into a new dataset of the same size as Y. The original locations of the observations of the bootstrapped data are stored for use before the maximization step to map the expectations from original data Y to the bootstrapped dataset. Each of the m bootstrapped datasets are constructed similarly using different random sequences and the entire bootstrapping step is completed before the EM loop depicted in Figure 1 begins. To improve the discrepancy among the set of m bootstrapped datasets we make use of quasi-Monte Carlo techniques introduced into bootstrapping methods (Teytaud et al., 2006; Kolenikov, 2007; Aidara, 2013). Specifically, we employ a scrambled (Matoušek, 1998) Sobol' sequence (Sobol', 1967) and conduct the bootstrapping following the steps outlined by Aidara (2013).

The following steps summarize the process for constructing the m bootstrapped datasets and m sets of location indicators used in MBEMMI:

Step 1: Create m column vectors of length N_i, each denoted as $\mathbf{x}_k = \{x_{1,k}, x_{2,k}, ..., x_{N_i,k}\}'$, where $k = \{1, 2, ..., m\}$ is the bootstrapped dataset indicator and $x_{i,k} \in \{0, 1, 2, ..., N_i\}$ is the number of times that the quarterly observation $\mathbf{y}_i$ is selected in bootstrapped dataset k.

Step 2: Generate m scrambled Sobol' sequences of length N_i and arrange them into the $N_i \times m$ matrix φ.

Step 3: Locate $\varphi_{1,1}$ and find $\inf\{x_{1,1} : \varphi_{1,1} \leq Prob(X_{1,1} \leq x_{1,1})\}$ and store the result as $x_{1,1}$, where $X_{1,1}$ has a binomial distribution with size N_i and probability $\frac{1}{N_i}$.

Step 4: For each $x_{i,1}$, where $i = \{2, 3, ..., N_i\}$, locate $\varphi_{i,1}$ and define $x_{i,1} = \inf\{x_{i,1} : \varphi_{i,1} \leq Prob(X_{i,1} \leq x_{i,1})\}$, where $X_{i,1}$ has binomial distribution with size $N_i - \sum_{l=1}^{i-1} x_{l,1}$ and probability $(1/N_i)/(1 - \frac{i-1}{N_i})$.

Step 5: Repeat Steps 3 and 4 $m - 1$ times to obtain the rest of the m frequency column vectors $\mathbf{x}_k$ for $k = \{1, 2, ..., m\}$.

Step 6: For the k^{th} bootstrapped dataset, select the y_i quarterly observation $x_{i,k}$ times and stack the selected quarterly observations into dataset Y_k', which has the same size as Y'. The rows of dataset Y_k' are randomly permuted since the order of the quarterly observations does not influence the expectation step. Repeat this selection process m times to generate m bootstrapped datasets.

D The MBEMMI and PSI Algorithms

Algorithm 1 MBEMMI

```
Input: $\mathbf{Y} = \{Y_1, Y_2, ..., Y_{N_j}\}$: list of series, some/all of which may have missing values.
       $\mathbf{a}$: list of vertical aggregations of $\mathbf{Y}$.
       $\mathbf{q}$: list of horizontal aggregations of $\mathbf{Y}$.
       $N_i$: number of observations in $\mathbf{Y}$.
       $N_j$: number of series in $\mathbf{Y}$.
       $m$: number of imputations.
       $tol$: convergence tolerance.
Output: $\mathbb{Y} = \{\mathbb{Y}_1, \mathbb{Y}_2, ..., \mathbb{Y}_m\}$: list of imputed $\mathbf{Y}$.
Process:
Step 1: Bootstrap $\mathbf{Y}$;
    save $m$ bootstrapped datasets as $\{\mathbf{Y}'_1, \mathbf{Y}'_2, ..., \mathbf{Y}'_m\}$
Step 2: Impute missing values;
    parfor $\mathbf{Y}'_k$ in $\{\mathbf{Y}'_1, \mathbf{Y}'_2, ..., \mathbf{Y}'_m\}$ do
        compute sufficient statistics $Q' = (\mathbf{Y}'_k)^T(\mathbf{Y}'_k)$
        $Q \leftarrow Q' * 0$
        while $||Q' - Q|| \geq tol$ do
            $Q \leftarrow Q'$
            (Expectation):
                estimate distribution of missing values, $(\widehat{\mu}, \widehat{\Sigma})$, from $Q$
                insert expectations of missing values, $\widehat{\mu}$, in original dataset $\mathbf{Y}$
            (Incorporating Aggregations):
                compute conditional distribution $((\widehat{\mu}'|\mathbf{a}, \mathbf{q}), (\widehat{\Sigma}'|\mathbf{a}, \mathbf{q}))$
            (Maximization):
                insert expectations of missing values, $\widehat{\mu}'$, in bootstrapped dataset $\mathbf{Y}'_k$
                compute sufficient statistics $Q'$
        end while
        compute converged conditional distribution $((\widehat{\mu}^*|\mathbf{a}, \mathbf{q}), (\widehat{\Sigma}^*|\mathbf{a}, \mathbf{q}))$
        draw one imputation for each missing value
        insert imputation in original dataset $\mathbf{Y}$, obtain imputed dataset $\mathbb{Y}_k$
    end parfor
return $\mathbb{Y} = \{\mathbb{Y}_1, \mathbb{Y}_2, ..., \mathbb{Y}_m\}$
```

Algorithm 2 PSI

```
Input: $\mathbf{D} = \{\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_L\}$: list of multi-level series.
       $\mathbf{D}_l = \{Y_{l,1}, Y_{l,2}, ..., Y_{l,N_l}\}$: list of series in level $l$, aggregations of series in $\mathbf{D}_{l+1}$.
       $\mathbf{a} = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_L\}$: vertical/temporal aggregations of $\mathbf{D}$.
       $\mathbf{a}_l = \{\mathbf{a}_{l,1}, \mathbf{a}_{l,2}, ..., \mathbf{a}_{l,N_l}\}$: vertical/temporal aggregations of $\mathbf{D}_l$.
       $L$: number of levels in multi-level data $\mathbf{D}$.
       $N_l$: number of series in level $l$.
       $m$: number of imputations.
Output: $\mathbb{D} = \{\mathbb{D}_1, \mathbb{D}_2, ..., \mathbb{D}_m\}$: list of imputed $\mathbf{D}$.
Process:
Step 1: Block $\mathbf{D}$ and $\mathbf{a}$;
    save blocks as $\mathbf{B}_{l,j} = \{Y_{l,j}, Y_{l+1,*}, \mathbf{a}_{l+1,*}\}$
Step 2: Impute missing values;
    parfor $k$ in $1 : m$ do
        for $l$ in $1 : (L-1)$ do
            parfor $\mathbf{B}_{l,j}$ in $\mathbf{B}_l = \{\mathbf{B}_{l,1}, \mathbf{B}_{l,2}, ..., \mathbf{B}_{l,N_l}\}$ do
                if $Y_{l+1,*}$ do not have missing values then
                    Continue
                else
                    if $l == (L-1)$ then
                        (MBEMMI) impute missing values in $Y_{l+1,*}$ once
                    else
                        (MBEMMI) estimate distributions of missing values, $(\widehat{\mu}, \widehat{\Sigma})$
                        insert expectations of missing values, $\widehat{\mu}$, in $\mathbf{B}_{l+1}$
                    end if
                end if
            end parfor
        end for
        for $l$ in $(L-2) : 1$ do
            compute missing values in $\mathbf{D}_l$ from linear constraints
        end for
        save imputed dataset as $\mathbb{D}_k$
    end parfor
return $\mathbb{D} = \{\mathbb{D}_1, \mathbb{D}_2, ..., \mathbb{D}_m\}$
```

E Summary of the Random Suppression Datasets

To validate the MBEMMI method, we use ten randomly suppressed Florida QCEW datasets. For each dataset, we first randomly suppress the quarterly employment counts in the fully-observed Florida QCEW data provided by Florida Department of Economic Opportunity (DEO). Then apply recursive secondary suppression (Cohen and Li, 2006) to protect the initial suppressions from being computed from the linear aggregations. As shown in Table 9, in the ten randomly suppressed datasets, there are on average 405.8 industries have suppressed values. Within those industries with suppressded values, the average suppression rate is 28.48%.

Table 9: Summary of the ten randomly suppressed Florida QCEW data. Industry count, incomplete industry count, and mean missing rate of the incomplete industries (95% CI) grouped by NAICS code levels.

Level	Industry Count	Incomplete Count	Incomplete Mean Missing %
2-digit	25	0.2 (0, 0.5)	10% (10%, 10%)
3-digit	94	3.4 (1.88, 4.92)	20.59% (17.27%, 23.91%)
4-digit	316	22.2 (19.73, 24.67)	26.15% (24.62%, 27.68%)
5-digit	679	111.3 (105.44, 117.16)	28.38% (27.6%, 29.16%)
6-digit	1043	268.7 (265.07, 272.33)	28.83% (28.33%, 29.33%)
Total	2157	405.8 (397.87, 413.73)	28.48% (28.08%, 28.89%)