Multiple Imputation for Large Hierarchical Multidimensional Data with Linear Aggregation Constraints

January 2, 2024

Abstract

Multiple imputation (MI) has became increasingly popular for handling missing data in empirical studies in recent years. However, current MI methods often struggle with accuracy and efficiency when applied to large hierarchical, multidimensional datasets subject to linear aggregation constraints, such as the Quarterly Census of Employment and Wage (QCEW). The inherent aggregations in such complex data structure pose significant challenges, as failing to account for them can often result in less accurate imputations and violations of linear aggregation constraints. This paper introduces a novel MI method designed to efficiently account for linear aggregation constraints. The method leverages singular normal distributions to address the aggregations while uses a Expectation Maximization (EM) algorithm with a Parallel-Sequential Imputation (PSI) scheme to handle large and complex datasets. Testing on real QCEW datasets demonstrates that the new method obtains twice the accuracy of leading alternatives while being ten times faster. Furthermore, an empirical application shows how the new method enables researchers to obtain unbiased estimates and make robust inferences.

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

1 Introduction

Missing data is a prevalent challenge in empirical economic studies. Missing data problem may arise from subject non-response, system failures, measurement errors, confidential suppressions or other causes. Since most analytical methods require complete datsets, researchers often have to either remove or fill the incomplete observations, or implicitly incorporate the missing data into their models, to facilitate meaningful analysis and inference. Contingent 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 lise-wise deletion, involving the exclusion of incomplete observations. However, this approach causes biased results if the data is 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 obtain biased results. Over recent decades, more sophisticated methods, such as maximum likelihood estimation, weighting, and various imputation methods have been developed for scenarios where the MCAR assumption does not hold. Among these, imputation methods have gained increasing popularity in recent years. Once the missing values are imputed, researchers can apply any analytical methods of interest and be free of missing data problem. Especially, the multiple imputation methods advance the other imputation methods because they can generate accurate imputations while adequately capture the uncertainty caused by 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). Collectively, these methods effectively impute common data types, including continuous, categorical, survival, longitudinal, and panel data

(Van Buuren, 2018; Little and Rubin, 2019).

For longitudinal or panel data studies, the most widely used multiple imputation methods among researchers 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 computational efficient and effective for large datasets. MCMC-based methods are robust for datasets with complex relationships and distributions.

However, existing multiple imputation methods struggle with large 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, undermining subsequent analyses. Moreover, neglecting these aggregations risks losing crucial information, potentially arise the Missing Not At Random (MNAR) problem and significantly reducing imputation quality.

Incorporating linear aggregations directly into MICE and EM-based methods is problematic as the aggregations introduce perfect colinearity. Several extensions of the MCMC-based methods are specialized to handle the linear constraints. 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 fall short in supporting multidimensional linear aggregation constraints. The Bayesian Multiscale Multiple Imputation (BMMI) method (Holan et al., 2010) uses singular normal distributions to model the linear aggregations into the MCMC process. But its MCMC process is computational intensive and needs a long time to ensure convergence, which is too slow for large datasets.

In this paper, we introduce a novel method, Multidimensional Bootstrapping Expectation Maximization Multiple Imputation method (MBEMMI), 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, it employs an EM algorithm that has deterministic convergence and incorporates a novel Parallel Sequential Imputation (PSI) scheme for easy parallelization.

Tests on real QCEW data and variants demonstrate that the MBEMMI method is two times more accurate than the leading MCMC alternative BMMI while maintaining competitive processing speed as the EM-based method EMB. MBEMMI takes about five minutes to generate ten imputed QCEW data set while EMB takes two and BMMI takes fifty. In a estimation of fixed effect model of average weekly wage, MBEMMI yielded unbiased point estimates and recovered standard errors comparable to a complete data scenario, while the complete case study obtained biased estimates and failed to make correct inference because the standard errors are too large.

The remainder of the paper is organized as follows: Section 2 dives into the data structure using a QCEW data sample; then Section 3 details the MBEMMI method, focusing on its approach to estimate the distribution of missing values; 4 adapts MBEMMI, BMMI, and EMB methods to the PSI scheme for large dataset handling, with validation on real QCEW datasets in Section 5; 6 tests the new method in model estimation, and Section 7 concludes the paper.

2 Data Structure

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

For illustration, Table 1 shows a sample of the disclosed 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 industry. In addition to the quarterly counts, the sample also contains annual aggregations (noted in every fifth row) and industry-wide totals (in 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 disclosed Florida QCEW data. The suppressed values are 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 have to incorporate the linear aggregation constraints when estimating the distribution of missing values. Successfully doing so not only make the imputations meet the constraints, but also potentially extracts additional information of the missing values from these constraints, enhancing imputation accuracy.

[1] The BLS does not explicitly disclose its suppression rule, but the 80/3 rule is widely accepted as the closest approximation (BLS, 2017).

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 multi-colinearity 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, MCMC process requires a 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 in speed, especially when the data set is large.

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 2,678 industries. 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%.

Table 2: Statistics 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	NaN
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, containing thousands of units, is extremely time consuming. The most efficient strategy is parallelization of the imputation processes. However, methods like BMMI requires some automatic convergence detection tools or long enough chains to ensure a convergence. In addition, priors that meet certain criteria need to be provided to each parallel job to ensure the imputations obtained in parallel can sufficiently capture the uncertainty. Although EMB method can be easily parallelized, it has challenges to account for linear aggregation constraints. Thus, currently there is no multiple imputation method can impute large hierarchical multidimensional data both accurately and swiftly.

3 Multidimensional Bootstrapping Expectation Maximization Multiple Imputation

In this section, we delve into the framework of the Multidimensional Bootstrapping Expectation Maximization multiple Imputation (MBEMMI) method. The discussion progresses from modeling data series, incorporating linear aggregations, to detailing how the multiple imputations are made.

MBEMMI begins with the assumption that, in the absence of linear aggregations, the data exhibit multivariate normally distribution (MVN) in at least one dimension, i.e. $Y \sim \mathcal{N}(\mu, \Sigma)$. It can be time dimension as in the QCEW example shown in Table 1, or geographic dimension as in a consumption pattern survey. The MVN assumption might seem strong in analysis, but it generally holds true for large data sets like QCEW. Also, researchers have shown that a MVN assumption usually works as well as the other complicated alternatives in multiple imputation practise (Schafer (1997), Schafer and Olsen (1998)).

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

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

where i denotes observation, j denotes variable, $\mathbf{y}_{i,-j}^{obs}$ are variables observed in i, and error term $\epsilon_i \sim \mathcal{N}(0, \nu^2)$. From 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 with the presence of missing values, the MBEMMI method uses a Expectation Maximisation (EM) process similar to Honaker and King (2010), which iteratively estimates the linear models using the sufficient statistics Q while replacing the missing values with the current expectations until convergence.[2]

Without considering the linear aggregation constraints, the EM process inevitably converges at a local maximum, where the estimates of the distributions of the missing values do not meet the linear aggregation constraints. Moreover, the estimates are less accurate as the extra information from the linear constraints is neglected.

To take the aggregations into account, in each iteration of the EM process, the MBEMMI method uses singular covariance matrices to correct the estimates of the missing data distribution and force them to meet the linear constraints.

Specifically, similar 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 the smallest unit that keeps the multidimensional linear constraint structure. Take the QCEW sample in Table 1 for example, one of the BCUs contains the first five rows, which are everything in year one. Anything smaller than that would break one or more linear constraints. The MBEMMI method transform the BCUs into vectors $z_{i'}$, where i' denote 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 $\Sigma_{i'}$ is singular as there is redundant information contained in the linear aggregations. Then MBEMMI partitions vector $z_{i'}$ into observed values $z_{i',o}$ and missing values $z_{i',m}$, the subsequent mean vector is then $\mu_{i'} = (\mu_{i',o}, \mu_{i',m})$, and covariance matrix:

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

Using the Moore-Penrose inverse $\Sigma^{+}_{i',oo}$ (Searle, 1982), the MBEMMI method corrects the

[2] Details of the EM process can be found in Appendix A

estimated distribution of missing values in each step of the EM process:[3]

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

With the converged distributions of missing values, the MBEMMI method can make random draws that satisfies the linear constraints. To fully explore the uncertainty of the missing values, instead of drawing multiple imputations from one converged distribution, similar to Honaker and King (2010), the MBEMMI uses a Quasi-Monte Carlo bootstrapping method to create m variations of the incomplete data set, and conduct independent EM processes to estimate m missing value distributions.[4] Then for each estimated distribution, MBEMMI makes one imputation. This bootstrapping-initiated multiple imputation method is faster than the maximum likelihood and IP methods (Honaker and King, 2010). In addition, it naturally supports "embarrassingly parallel" as the bootstrapped processes are independent. Moreover, the EM processes converge deterministically, which means no expert supervision are needed.

As shown in Figure 1, the MBEMMI method consists of the following key steps:[5]

Step 1: Bootstrap data sets $(\mathbf{Y}'_1, \mathbf{Y}'_2, \ldots, \mathbf{Y}'_m)$. For each bootstrapped data set $\mathbf{Y}'_k$, apply step 2-6.

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

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

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

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

[3] Detailed steps of incorporating multidimensional linear aggregations can be found in Appendix B.

[4] Detailed Quasi-Monte Carlo bootstrapping method can be found in Appendix C.

[5] Also see Algorithm 1 in Appendix D

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

4 Parallel Sequential Imputation Scheme

As the multiple imputation processes are independent and can automatically converge, the MBEMMI method can be easily scaled for large hierarchical multidimensional data through a Parallel Sequential Imputation (PSI) blocking scheme.

Taking the 2012-2016 Florida QCEW data for example (Table 2), imputing $2,157$ industries while accounting for the linear aggregations is not only extremely time consuming, but may also crash the process while inverting huge covariance matrices. Since the QCEW data uses a NAICS code structure, which is tree-like and has five hierarchical levels, we can separate it into small blocks, each only contains one coarser resolution industry and all its immediate sub-industries, just like the QCEW example in Table 1. As the highly correlated industries are already grouped by NAICS codes under the same coarser resolution industry, and researchers found that most inter-block correlation is already captured by the linear aggregations (Holan et al., 2010), we can treat the small blocks as they are independent. Thus, they can be imputed separately, which makes the data more manageable.

Also from Table 2, we notice the higher (coarser resolution) levels tend to have fewer incomplete industries. That is because as the level increases, the vulnerable industries are gradually covered in the aggregations, therefore, less suppression is needed. This is beneficial to imputation purposes as the information of the missing values are still in the higher level aggregations, only it is mixed together. By imputing the small blocks top-down level-by-level, we carry the information of missing values to the lower levels, where it is needed the most. This level-by-level scheme brought more information to the imputation and helps it yield better results.

In the PSI scheme, we impute the QCEW data level-by-level sequentially, while in each level, the separated small blocks are imputed in parallel. When finished, we make one imputation for the whole data set, can repeat the scheme m times in parallel to obtain m

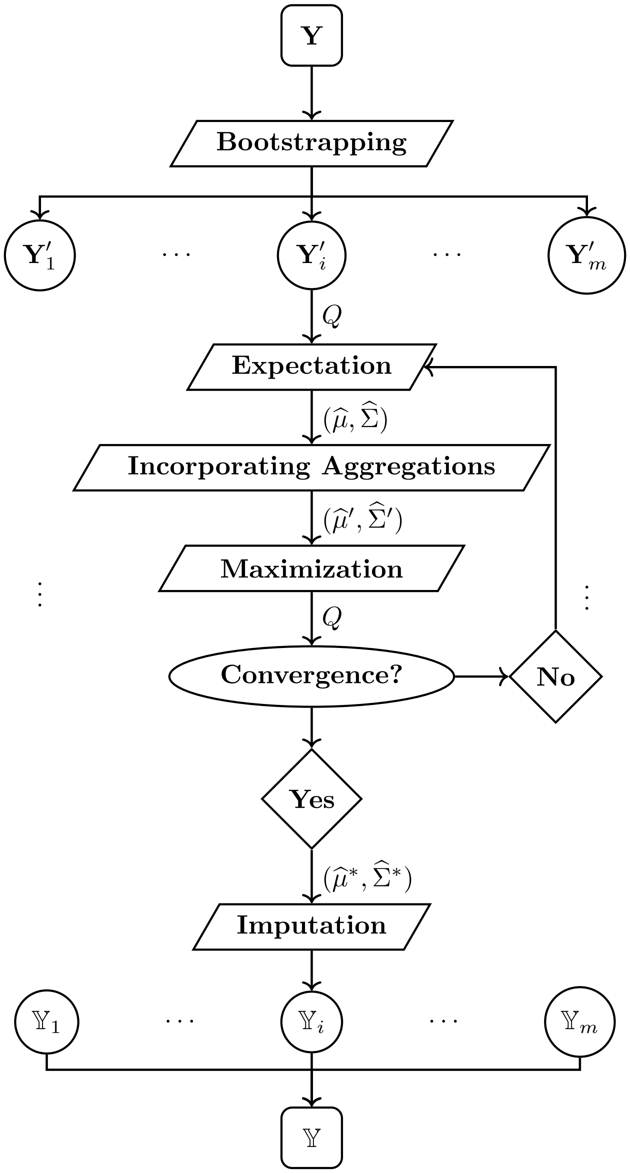

Figure 1: The MBEMMI Method

imputations.[6]

The MBEMMI method is the only multiple imputation method that fully supports the PSI scheme. The BMMI method needs automatic tools to detect convergence. Moreover, certain prior rules have to be employed to ensure the quality of the imputations obtained from parallel processes in stead of a single Markov chain. The EMB method can not account for the linear aggregations, which on one hand weakens the assumption of inter-block Independence, on the other hand makes the imputation less accurate.

In the next section, we compare the performances of the fore-mentioned multiple imputation methods on real QCEW data.

5 Validating the Algorithm

This section validates the MBEMMI method in terms of accuracy and speed, comparing it against two leading alternatives, BMMI and EMB, using real QCEW data. We start with the QCEW sample from Table 1 for a detailed examination, then move on to large QCEW datasets with the help of the PSI scheme and report the performances statistics.

The MBEMMI-imputed QCEW sample is shown in Table 3. Each missing quarterly employment count was imputed $m = 10$ times. Subsequently, we obtain ten completed data sets. The missing annual totals for series 1 and 2 in year 2 can be computed by adding the year 2 quarterly counts together. In the completed data sets, all linear aggregation constraint hold.

Since the missing values in the QCEW sample are suppressed for confidential reasons, I can not directly display the imputed values nor compare them with the true values. Instead, I display the 95% confidence intervals in bold numbers in the missing cells. Notice that the intervals generally align with the observed values immediately before or after the missing values, indicating the imputations can preserve the trend of the employment time series and the imputed values do not severely deviate from the observed values. The widths of the confidence intervals depend on the uncertainty caused by the missing values. For instance, the intervals in year 2 tend to be wider than those in year 3 as there are more missing values in year 2, even two annual totals are also missing. More missing values means more

[6] The detailed steps of the PSI scheme is in Algorithm 2 in Appendix D.

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

information loss, thus, higher uncertainty when making imputations.

For confidentiality reasons, the true values for the suppressed data in the QCEW sample cannot be displayed. As an alternative, we replace all values in Table 3 with data from industries which are not vulnerable to intruders and hence not suppressed. The exact same suppression from Table 3 is applied. With the new data, we can explicitly validate the methods as the ”true” values in it are not confidential.

Imputations for the new QCEW sample were generated using MBEMMI, BMMI, and EMB, each producing $m = 10$ imputations. Figure 2 shows the full series 1 and 2, and the confidence intervals of the imputed values as error bars. The error bars in the top figure appear smaller than those in the bottom figure, but numerically they are similar in size. It

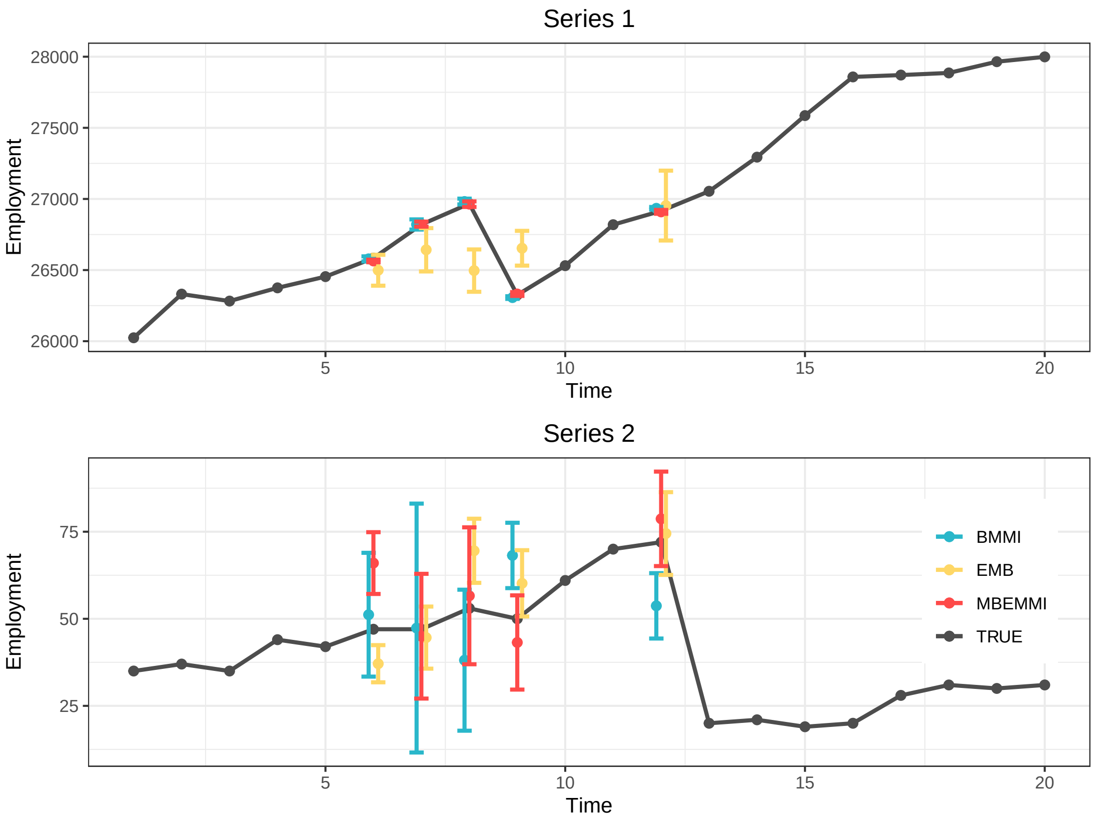

Figure 2: Imputed QCEW sample in Table 1. For confidentiality reasons, the values in the sample are replaced with disclosed data while the suppressions are unchanged.

is because the imputations are constrained by the industry-wise aggregations, which means if a imputation in series 1 at time i, i.e. $\tilde{y}_{i,1}$, increases by s, to hold the linear aggregation, the corresponding imputation $\tilde{y}_{i,2}$ must decrease s. This property would lead to relatively wider error bars in smaller industry (industry 2) while smaller error bars in larger industry (industry 1).

In series 1, both MBEMMI and BMMI show high accuracy, while EMB exhibits larger error bars and three of them missed the true values. In series 2, all three methods work similarly in accuracy, with MBEMMI slightly outperforming the others. MBEMMI's error bars missed one true value, BMMI missed two, and EMB missed three. The reason why EMB worked apparently better in series 2 than in series 1 is that it tends to assume the missing values follow existing trends, which is true in series 2 but not the case in series 1. Therefore, the EMB methods works better in situations where the series do not vary very much. The MBEMMI and BMMI method on the other hand, work very well regardless of the trends as they borrow strength from the linear aggregation constraints.

We now validate the methods on more QCEW data. As discussed in section 4, we use PSI scheme to separate large hierarchical multidimensional data sets into small blocks to avoid imputing the whole data set altogether. The individual blocks are imputed in a parallel-sequential manner that increases both the imputation speed and the resistance to failures caused by large data sets.

To impute the whole Florida QCEW data set, we first use PSI scheme to separate the large data set into small blocks (Table 4). Out of 1,114 blocks, 69 of them have missing values and need to be imputed. However, most of them contain series of severely missing values, i.e. series have more than 60% values missing. In fact, 28 of them contain completely missing series, 8 contain series missing more than 80% values. As the severely missing series contain little information for meaningful imputations (Rubin, 1996), in this study we focus on the 28 blocks that have missing rates less than 60%.

Table 4: Blocks in 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 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

Using MBEMMI, BMMI, and EMB, we imputed these 28 blocks, each method makes $m = 10$ imputations. Then the imputations are compared to the true values. I use a metric called "τ% hit-rate" to measure the accuracy of the MI methods. The metric means the percentage of imputed values are within τ% of the true values:

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

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

The hit-rates ψ_1^p, ψ_2^p, ψ_5^p, and ψ_{10}^p for each method on the 28 QCEW samples are shown in Table 5. In every category, the MBEMMI method hits more targets than the other two methods. Especially, 10.58% of the MBEMMI imputations are within the 1% interval of the true values, which more than doubled the BMMI method, more than tripled the EMB method. The BMMI method has better hit-rates than the EMB method as it can account for the linear aggregation constraints while EMB can not.

Table 5: Percentage of imputed values within 1%, 2%, 5%, 10% of the true values. The QCEW samples are separated from the Florida QCEW data set 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%

In addition to accuracy metrics, Table 6 also shows the average speed of each method. We can see the EMB method uses on average 0.02 second to produce one imputed QCEW sample. The MBEMMI method uses around 0.63 second, which is the second fastest. The BMMI method uses the longest as a 8,000 burn in period is necessary for the Markov chains to converge.

Table 6: Average speed and 95% confidence intervals per one QCEW sample imputation. All tests are in R. 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. Platform: Apple MacBook M1 Pro, 8 cores, 16 GB 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)**

To explore the full potential of the MBEMMI method in imputing large hierarchical multidimensional data sets, we use 10 randomly suppressed QCEW data sets. For each data set, I randomly suppress the fully-observed confidential Florida QCEW data set, then conduct recursive secondary suppression (Cohen and Li, 2006) to protect the initial suppressions from being computed from the linear aggregations. The random suppression data sets do not have severely missing series,[7] which means we do not need to exclude the problematic series and break the NAICS structure. We can focus on how the MI methods work on the entire large data sets.

Following the PSI scheme, each random suppression data set is imputed $m = 10$ times in a parallel-sequential manner. I show the pooled hit-rates of each MI method in Table 7. We can notice all methods have better hits than in the QCEW samples. It is because the missing rates are lower in the random suppression data sets. The MBEMMI method has the highest hit rates in all categories. Especially, in the high accuracy categories ψ_1 and ψ_2, MBEMMI performs twice as good 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.

Table 7: Percentage of imputed values within 1%, 2%, 5%, 10% of the true values. The ten randomly suppressed Florida QCEW data sets 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 can also find the average speeds of the MI methods in Table 8. The EMB method uses around 0.01 second to impute a single block once, while the MBEMMI method takes 0.29 seconds. Both methods are much faster than the BMMI method. The random suppression data sets also enable us to test the speeds of the MI methods on full data sets. To create ten imputed QCEW data sets, the EMB method takes 1.85 minutes. The MBEMMI method uses on average 4.79 minutes. Both methods have practical speeds. On the contrary, the BMMI method needs around 51.23 minutes to finish.

[7] Please see Table 9 in Appendix E for a summary of the random suppression data sets.

Table 8: Average speed and 95% confidence intervals per one random suppression block imputation and per ten random suppression data set imputations. All tests are in R. 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 data sets. Platform: Apple MacBook M1 Pro, 8 cores, 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)**

The tests on samples and full size of QCEW data set demonstrate that the new multiple imputation method MBEMMI can impute large hierarchical multidimensional data set both accurately and fast in speed. The next section explores MBEMMI's application in a panel data study of wage effects, highlighting how researchers can benefit from applying MBEMMI instead of using only complete cases in their study.

6 Empirical Application: Average Weekly Wage

Now we apply the MBEMMI method to a panel data model analyzing average weekly wages. The model is specified as follows:

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

Here, we explore the effects of employment level, establishment count, and macroeconomic indicators (GDP growth rates, inflation rates, and unemployment rates) on the average weekly wage levels in various industries. In this model, i' denotes year, j represents industry, $\upsilon_{i'}$ and φ_j are fixed time and unit effects. We assume that all classical assumptions about $\epsilon_{i',j}$ apply.

To illustrate the efficacy of MBEMMI, we first generate $m = 10$ imputed QCEW sample from Table 1 using MBEMMI, then estimate the fixed effect model 4 on every imputed data set and obtain 10 estimates, finally pool the results using Rubin's rules (Rubin, 1987)

The point estimates of the coefficients and their 95% confidence intervals are shown in Figure 3. On the contrary to the complete case results (blue), the MBEMMI results (red) are very similar to the truth (black). The complete case results have larger standard errors and biased point estimates, while the MBEMMI results are unbiased and correctly recover the standard errors. Consequently, the MBEMMI results successfully rejects the null hypothesis $\beta = 0$ for variables *Establishment*, *GDP Growth*, and *Unemployment* at 95% confidence, a task the complete case study fails to achieve.

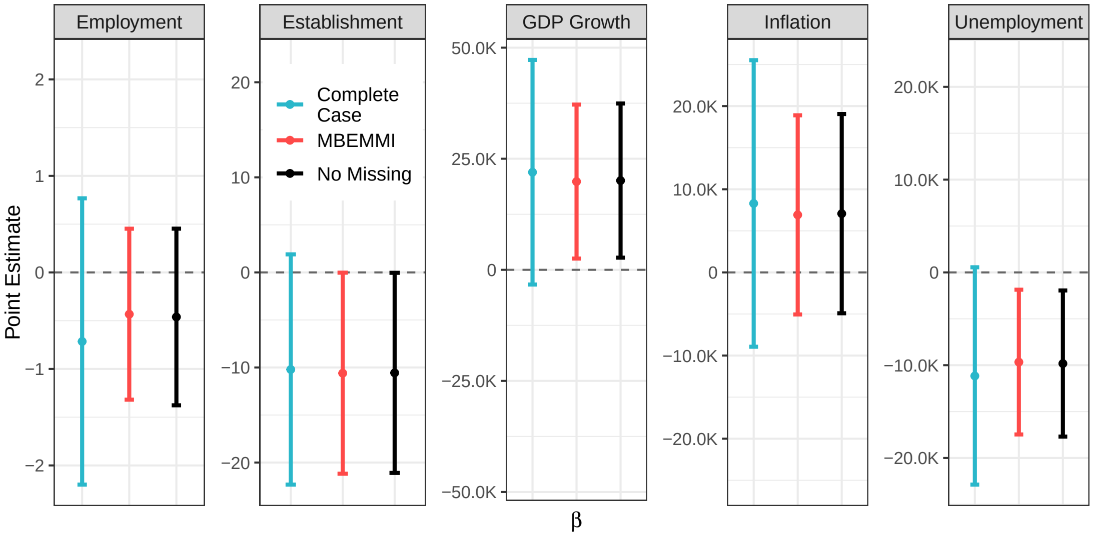

Figure 3: Point estimate and 95% confidence intervals. The black intervals are from original data with no missing values, blue intervals are complete case estimation of the suppressed data, red intervals are MBEMMI imputed data.

This application showcases how MBEMMI can assist researchers in obtaining estimates that closely resemble those derived from the complete data analysis with no missing values, and subsequently results in more accurate and reliable statistical inferences.

7 Concluding Remarks

In this study, we tackle the difficulties in imputing large hierarchical multidimensional data with linear aggregation constraints. The existing multiple imputation methods are either not accurate due to fail to take linear aggregations into account, or not fast enough to be a part of the seamless streamline in empirical research. As the data sets that Economists work on become increasingly complex and large, the multiple imputation methods have to be to compatible to the structure while can handle large data files in short amount of time.

To meet this need, we developed the Multidimensional Bootstrapping Expectation Maximization Multiple Imputation (MBEMMI) method, which employs singular normal distributions to account for the multidimensional linear constraint structure while uses EM algorithm along with a Parallel-Sequential Imputation (PSI) scheme to facilitate rapid imputation of large datasets. The new method ENABLES researchers to save considerable time in imputing large data sets while achieving comparable or superior accuracy compared to existing methods.

Through real-word data sets and an empirical application, we demonstrate that the MBEMMI method outperforms the leading alternatives in both accuracy and speed. It is two times more accurate in imputing the Florida QCEW samples and the randomly suppressed data sets. It is more than ten times faster than the only multiple imputation method that can take the linear aggregations into account. The model estimation also show the MBEMMI method helps researcher obtain unbiased estimates and make correct inferences.

Future research will be needed to explore the applications of MBEMMI on large data sets such as monthly GDP data, consumption survey response. Additionally, relaxed distribution assumptions will be explored and a versatile R package will be developed to broaden MBEMMI's usability and accessibility. The future work will further cement MBEMMI's role as a vital tool of researchers in dealing with complex and large 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 data set Y with their conditional expectations, based on the current estimates of the sufficient statistics of bootstrapped data.

The sufficient statistics are $Q = (Y_i')^T(Y_i')$, where Y_i' is the ith bootstrapped data set whose aggregation values are excluded to avoid perfect multi-collinearity, and the first column of Y_i' is a vector of ones. For instance, let Y_A' be the bootstrapped data set A, then Y_A' will be:

$$Y_A' = \begin{pmatrix} 1 & y_{4,1} & S & \cdots & S & B_4 \\ 1 & S & y_{7,2} & \cdots & S & B_7 \\ 1 & S & S & \cdots & y_{3,N} & B_3 \\ 1 & y_{2,1} & y_{2,2} & \cdots & y_{2,N} & B_2 \\ 1 & S & y_{7,2} & \cdots & S & B_7 \\ 1 & S & S & \cdots & y_{3,N} & B_3 \\ 1 & y_{5,1} & y_{5,2} & \cdots & y_{5,N} & B_5 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \end{pmatrix}$$

where, for instance, the first subscript of 4 in $y_{4,1}$ denotes that the first row of bootstrapped data set Y_A' was from quarter $t = 4$ from the original data set Y'. Recall that B_4 denotes the auxiliary variables which are always observed. The sufficient statistics for data Y_A' are computed as $Q_A = (Y'_A)^T(Y'_A)$. It is convenient to rewrite the sufficient statistics as

$$Q^*_A = \begin{pmatrix} -1 & \hat{\mu}_1 & \hat{\mu}_2 & \cdots & \hat{\mu}_N & \hat{\mu}_B \\ \hat{\mu}_1 & \hat{\sigma}^2_1 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1N} & \hat{\sigma}_{1B} \\ \hat{\mu}_2 & \hat{\sigma}_{12} & \hat{\sigma}^2_2 & \cdots & \hat{\sigma}_{2N} & \hat{\sigma}_{2B} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\mu}_N & \hat{\sigma}_{1N} & \hat{\sigma}_{2N} & \cdots & \hat{\sigma}^2_N & \hat{\sigma}_{NB} \\ \hat{\mu}_B & \hat{\sigma}_{1B} & \hat{\sigma}_{2B} & \cdots & \hat{\sigma}_{NB} & \hat{\sigma}^2_B \end{pmatrix},$$

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

The MBEMMI algorithm executes estimations by using a SWEEP operator $\theta(s)$ (Beaton, 1964). Given an $1 \times (1 + N + P)$ input vector s that consists of only 1's and 0's, where $(1 + N + P)$ is the number of columns in Q^*_A and P is the number of auxiliary variables, a SWEEP operator $\theta(s)$ will operate on the elements of Q^*_A and obtain corresponding estimates. The intuition of the SWEEP operator is to transform the Q^*_A 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

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

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

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

Note that $s = \{0, 1, 0\}$ means we are estimating $Y_2 = \beta_0 + \beta_1 Y_1 + \epsilon$. The values obtained in the third column above are the coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ and the variance $\hat{\sigma}^2_{2|1}$.

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

$$E(y_{t,n}) = y_t^{obs}\theta\{1 - M_t\}_n^i,$$

where $s = (1 - M_t)$ is the observed value indicator of quarter t, y_t^{obs} is the $1 \times (1 + N + P)$ vector of observations at time t with zeros in the missing cells, and $(1 + N + P) \times 1$ vector $\theta\{1 - M_t\}_n^i$ denotes that we are using the n^{th} column of swept matrix θ obtained in the i^{th} iteration. The transformation $Q \to Q^*$ 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 $E(y_{t,n}) = \beta_0 + \beta_{-n} \times y_{t,-n}$, where $-n$ stands for other numbers in $\{1, 2, 3\}$ but not n. The variance of the missing values are obtained as

$$\hat{\sigma}^2_{2|1} = \theta\{1 - M_t\}_{t,n}^i$$

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

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

For each observation y_t in original data set Y', the filled-in row vector $\hat{y}_t^E$ and covariance matrix of missing values $\Sigma^E_{t|y_t^{obs}}$ will be

$$\begin{aligned}\hat{y}_t^E &= y_t^{obs} + M_t \cdot (y_t^{obs}\theta\{1 - M_t\}^i)\\ \Sigma^E_{t|y_t^{obs}} &= M_t' M_t \cdot \theta\{1 - M_t\}^i\end{aligned}$$

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

Maximization Step

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

$$Q' = \sum_t (\hat{y'}_t^E \hat{y}_t^E + \Sigma^E_{t|y_t^{obs}}),$$

The sufficient statistics Q' will be used in the next E-Step to generate new expectations and variances for missing values of the original data set 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 $Q_A(i)$ from bootstrapped data set $Y'_A(i)$, which, recall, was previously stripped of the aggregation constraints in the raw data file Y, and substituted back into the appropriate locations in Y'. The multiscale updating step then enforces the binding constraints to produce and updated data set, say, $\tilde{Y}'(i)$. We then use the stored mapping vector to map the suppressed values back into an updated bootstrapped data set, $Y'_A(i+1)$. 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 multiscale updating step we reimpose the multiscale linear constraints using a technique based on the multiscale simulation step of Holan et al. (2010) and construct a new sufficient statistics matrix that incorporates these constraints.

The M-U step starts by transforming quarterly data y_t, quarterly aggregations q_t and annual aggregations a_n into a yearly vector

$$\begin{aligned} z_{t'} = (&y_{4t'-3,1}, ..., y_{4t',1}, y_{4t'-3,2}, ..., y_{4t',1}, ..., y_{4t'-3,N}, ..., y_{4,N}, \\ &q_{4t'-3}, ..., q_{4t'}, a_{t',1}, ..., a_{t',N})' \end{aligned}$$

where $t' = \{1, 2, 3, ..., \frac{T}{4}\}$ represents the index for years.

Using year 1 as an example, the transformation places the values of the first five rows of data set Y into a column vector z_1. z_1 uses only observable aggregation values. For example, note that q_3, q_4 and $a_{1,1}, a_{1,N}$ are not included in z_1 since they are unknown and cannot provide any information to infer the missing values $y_{3,1}, y_{3,2}, y_{4,2}$ and $y_{4,N}$. The element $a_{t,N-1}$ in vector z_1 is the last available industry annual total in year 1. The element a_1 aggregate annual total in year 1 is redundant and so is not included in z_1.

Define the operator matrix H:

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

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 identity matrices of size 4×4. For z_1, since the aggregation values are not complete, the corresponding operator matrix is

$$H_1 = \begin{pmatrix} & & & I_{4N} & & & \\ I_2 & 0_2 & I_2 & 0_2 & \cdots & I_2 & 0_2 \\ & & & I_{(N-2)\times N} \otimes 1_4' & & & \end{pmatrix},$$

where 0_2 is 2-by-2 matrix with only zeros and $I_{(N-2)\times N}$ is a (N-2)-by-N matrix transformed from identity matrix I_N by deleting the corresponding row of I_N whenever the annual total is not observed. In this example, these are rows 1 and N.

Given the appropriate operator matrix H, we obtain the mean vector $\mu_{t'}$ and covariance matrix Σ of $z_{t'}$ as

$$\begin{aligned} \mu_{t'} &= H\theta_{t'} \\ \Sigma &= HVH', \end{aligned}$$

where

$$\begin{aligned} \theta_{t'} &= (E(y_{4t'-3,1}), ..., E(y_{4t',1}), E(y_{4t'-3,2}), ..., E(y_{4t',2}), ..., E(y_{4t'-3,N}), ..., E(y_{4t',N})), \\ V &= diag(\hat{\sigma}_1^{2^E}, \hat{\sigma}_1^{2^E}, \hat{\sigma}_1^{2^E}, \hat{\sigma}_1^{2^E}, \hat{\sigma}_2^{2^E}, \hat{\sigma}_2^{2^E}, \hat{\sigma}_2^{2^E}, \hat{\sigma}_2^{2^E}, ..., \hat{\sigma}_N^{2^E}, \hat{\sigma}_N^{2^E}, \hat{\sigma}_N^{2^E}, \hat{\sigma}_N^{2^E}). \end{aligned}$$

In $\theta_{t'}$, 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 $\hat{\sigma}_{nk}^E$ 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 $z_{t',m}$ conditional on observed values $z_{t',o}$ is

$$z_{t',m} \mid z_{t',o} \sim N(\gamma_{t',m}, \Omega_m),$$

where the mean vector $\gamma_{t',m}$ and covariance matrix Ω_m are given by

$$\begin{aligned} \gamma_{t',m} &= \mu_{t',m} + \Sigma_{mo}\Sigma_{oo}^{+}(z_{t',o} - \mu_{t',o}), \\ \Omega_m &= \Sigma_{mm} - \Sigma_{mo}\Sigma_{oo}^{+}\Sigma_{om}. \end{aligned}$$

At the completion of the M-U step the posterior distributions of the missing variables based upon a particular bootstrapped data set 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 1 we first eliminate the annual total rows ($a_{t,i}$'s) and the quarterly totals (q_t's) to get a $T \times N$ array of $y_{t,i}$'s. Next, to each row we add a vector of auxiliary variables, B_t, 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 data set 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 multiscale updating step so they are kept aside in original data set Y along with mapping indicators to allow us map the aggregations in Y back to the detailed values in Y' in the multiscale updating (M-U) step.

The bootstrapping process randomly picks one quarterly observation at a time with replacement from Y' and stacks them into a new data set 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 data set. Each of the m bootstrapped data sets are constructed similarly using different random sequences and the entire bootstrapping step is completed before the multiscale 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 recently 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 data sets and m sets of location indicators used in MBEMMI:

Step 1: Create m column vectors of length T, each denoted as $B_h = \{x_{1,h}, x_{2,h}, ..., x_{T,h}\}'$, where $h = \{1, 2, ..., m\}$ is the bootstrapped data set indicator and $x_{t,h} \in \{0, 1, 2, ..., T\}$ is the number of times that the quarterly observation y_t is selected in bootstrapped data set h.

Step 2: Generate m scrambled Sobol' sequences of length T and arrange them into the $T \times m$ 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 T and probability $\frac{1}{T}$.

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

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

Step 6: For the h^{th} bootstrapped data set, select the y_t quarterly observation $x_{t,h}$ times and stack the selected quarterly observations into data set Y_h', which has the same size as Y'. The rows of data set Y_h' 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 data sets.

D The MBEMMI and PSI Algorithms

Algorithm 1 MBEMMI

```
Input: $\mathbf{Y} = \{Y_1, Y_2, ..., Y_N\}$: list of time series, some/all of which may have missing values.
       $A = [a_{1:1}, a_{1:2}, ..., a_{T:K}]$: vector of logical/spatial aggregations of $\mathbf{Y}$.
       $B = \{B_1, B_2, ..., B_N\}$: list of temporal aggregations of $\mathbf{Y}$.
       $N$: number of time series in $\mathbf{Y}$.
       $T$: number of temporal aggregations in $B_j$.
       $K$: length of $Y_i$ in each temporal aggregation.
       $m$: 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 data sets as $\{\mathbf{Y}'_1, \mathbf{Y}'_2, ..., \mathbf{Y}'_m\}$
Step 2: Impute missing values;
    parfor $\mathbf{Y}'_i$ in $\{\mathbf{Y}'_1, \mathbf{Y}'_2, ..., \mathbf{Y}'_m\}$ do
        compute sufficient statistics $Q'$
        $Q \leftarrow Q' * 0$
        while $||Q' - Q|| \geq tol$ do
            $Q \leftarrow Q'$
            (Expectation):
                estimate distribution of missing values, $(\bar{\mu}, \bar{\Sigma})$, from $Q$
                insert expectations of missing values, $\bar{\mu}$, in original data set $\mathbf{Y}$
            (Multiscale-Correction):
                compute conditional distribution $((\bar{\mu}'|A, B), (\bar{\Sigma}'|A, B))$
            (Maximization):
                insert expectations of missing values, $\bar{\mu}'$, in bootstrapped data set $\mathbf{Y}'_i$
                compute sufficient statistics $Q'$
        end while
        compute converged conditional distribution $((\bar{\mu}^*|A, B), (\bar{\Sigma}^*|A, B))$
        draw one imputation for each missing value
        insert imputation in original data set $\mathbf{Y}$, obtain imputed data set $\mathbb{Y}_i$
    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 time series.
       $\mathbf{D}_l = \{Y_{l,1}, Y_{l,2}, ..., Y_{l,N(l)}\}$: list of time series in level $l$, aggregations of $\mathbf{D}_{l+1}$.
       $\mathbf{B} = \{\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_L\}$: temporal aggregations of $\mathbf{D}$.
       $\mathbf{B}_l = \{B_{l,1}, B_{l,2}, ..., B_{l,N(l)}\}$: temporal aggregations of $\mathbf{D}_l$.
       $L$: number of levels in multi-level data.
       $N(l)$: number of time 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: Partition $\mathbf{D}$ and $\mathbf{B}$;
    save partitions as $\mathbf{P}_{l,j} = \{Y_{l,j}, Y_{l+1,*}, B_{l+1,*}\}$
Step 2: Impute missing values;
    parfor $i$ in $1 : m$ do
        for $l$ in $1 : (L-1)$ do
            parfor $\mathbf{P}_{l,j}$ in $\mathbf{P}_l = \{\mathbf{P}_{l,1}, \mathbf{P}_{l,2}, ..., \mathbf{P}_{l,N(l)}\}$ 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, $(\bar{\mu}, \bar{\Sigma})$
                        insert expectations of missing values, $\bar{\mu}$, in $\mathbf{P}_{l+1}$
                    end if
                end if
            end parfor
        end for
        for $l$ in $(L-2) : 1$ do
            compute missing values in $Y_l$ from linear constraints
        end for
        save imputed data set as $\mathbb{D}_i$
    end parfor
return $\mathbb{D} = \{\mathbb{D}_1, \mathbb{D}_2, ..., \mathbb{D}_m\}$
```

E Statistics of the Random Suppression Data Sets

Table 9: Statistics 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	25.00	0.2 (0, 0.5)	10% (10%, 10%)
3-digit	94.00	3.4 (1.88, 4.92)	20.59% (17.27%, 23.91%)
4-digit	316.00	22.2 (19.73, 24.67)	26.15% (24.62%, 27.68%)
5-digit	679.00	111.3 (105.44, 117.16)	28.38% (27.6%, 29.16%)
6-digit	1043.00	268.7 (265.07, 272.33)	28.83% (28.33%, 29.33%)