Abstract. Missing time-series data is a prevalent practical problem. Imputation methods in time-series data often are applied to the full panel data with the purpose of training a model for a downstream out-of-sample task. For example, in finance, imputation of missing returns may be applied prior to training a portfolio optimization model. Unfortunately, this practice may result in a look-ahead-bias in the future performance on the downstream task. There is an inherent trade-off between the look-ahead-bias of using the full data set for imputation and the larger variance in the imputation from using only the training data. By connecting layers of information revealed in time, we propose a Bayesian posterior consensus distribution which optimally controls the variance and look-ahead-bias trade-off in the imputation. We demonstrate the benefit of our methodology both in synthetic and real financial data.

1. Introduction

Missing data is a common practical problem often encountered in data collection and, as such, is prevalent in virtually any area touched by data science. There is a large body of literature which deals with general imputation tools, see for example, [23]. Understanding the implications of imputing time-series data in downstream out-of-sample tasks, however, is significantly less studied. This is precisely the goal of this paper.

We propose a novel Bayesian approach which interpolates between the use of globally observed data (i.e. data collected during the whole time horizon) and locally observed data (i.e. data collected up to a fixed time in the past) in order to optimize the trade-off between look-ahead-bias (which increases when using globally observed data to impute) and variance (which increases when using only partial data to impute).

Naturally, the philosophy applied in virtually every imputation method involves extracting as much information as possible from the globally observed data to infer the missing data. However, in time-series settings (and others in which a downstream task occurs relative to some ordering), this philosophy poses a problem. This is of particular importance in financial applications with portfolio optimization as a leading example. Portfolio selection procedures are designed to exploit subtle signals that can lead to favorable future investment outcomes. As spurious signal may induce a poor strategy, the investments are typically evaluated out-of-sample to avoid potential overfitting by “looking into the future”.

However, missing data is a prevalent problem in financial data sets: A large number of instruments is traded over the counter or infrequently and many assets might only temporarily trade on public exchanges. If missing data is imputed globally using all information, an out-of-sample portfolio selection procedure on the imputed data will inherently suffer from a look-ahead-bias. This look-ahead-bias can be avoided by only using past data for imputation and hence avoid “looking into the future”. This means that sequentially more data is added to imputation procedure. While this approach avoids the look-ahead-bias problem, it ignores valuable information that could be used to improve the error in the imputation procedure. While overfitting by using future information has been studied in the finance literature [12, 8, 15], none of the existing results so far explores how to formulate and quantify look-ahead-bias and optimally trade off between look-ahead-bias and variance in imputations for future out-of-sample tasks.

Our conceptual idea extends to general imputation in time-series that are followed by a downstream task. We choose the finance setting as a particularly relevant application. While a look-ahead-bias in the financial context is generally understood as using data that is available only in the future, as for example in [8], the
exact quantification and definition is a challenging problem. We adopt a Bayesian imputation framework in order to precisely define look-ahead-bias as deviation from a specific baseline posterior distribution. We assume a Bayesian model with the observed entries as realizations. Such a model is endowed with a prior distribution for the underlying parameters. The full Bayesian model generates a corresponding posterior both for the parameters and the missing observations conditional on the observed data. This type of formulation is standard in the multiple imputation literature \cite{28,23}. The multiple imputation framework, from which we borrow inspiration, imputes by sampling (multiple times) from the posterior, creating multiple versions of imputed data set, each of which can be used for relevant downstream tasks.

The full data imputation (with look-ahead-bias and small variance) and conservative imputation (without look-ahead-bias and with large variance) are both special cases of the Bayesian approach. Each of them corresponds to a different posterior. The posterior computed by fitting the Bayesian model only on the training set (which we call baseline) is assumed to be look-ahead-bias-free, and the look-ahead-bias incurred by any other posterior is defined as the bias relative to the baseline posterior. Our approach differs from the standard multiple imputation literature in that we are interested in combining these two natural posteriors to minimize the bias and variance trade-off in the estimation of the returns. We therefore will obtain a “consensus” posterior that is optimal in the sense of this trade-off. Most of the literature in multiple imputation, in fact, would naturally advocate using the full panel data.

We illustrate our novel conceptual framework in an i.i.d. Gaussian model with a known covariance matrix. Thus, the moments of the posterior of missing entries reduce to the moments of the posterior of the mean of the Gaussian model. We use a variational approach to synthesize the two posteriors (fitted from the training set and the full panel) of the mean parameter. A single consensus posterior is selected by computing a Wasserstein barycenter interpolation of the two posteriors. The barycenter interpolation introduces weights which prescribe the importance attached to the mutual information between the consensus and each of the posteriors. This set of weights is further optimized to find the optimal consensus posterior according to a criterion that minimizes the variance of the consensus posterior while constraining its look-ahead-bias not to exceed some threshold. The threshold parameter can then be either chosen by cross validation or by minimizing the estimated mean squared error over all consensus posteriors as a function of the threshold parameter.

Finally, we empirically validate the benefit of our methodology with the downstream task of portfolio optimization, in which we show that the use of an optimal consensus posterior results in better performance of the regret of out-of-sample testing returns measured by what we call “expected conditional mean squared error”, compared to the use of natural posteriors. We demonstrate our method on both synthetic and real financial data, and under different practically relevant missing patterns, which include the simple case of missing at random, a dependent block missing structure, and when the missing pattern depends on past realizations of returns.

Contributions. In summary, our contributions are

- We propose a novel conceptual framework to formalize and quantify the look-ahead-bias in the context of data imputation. We combine the different layers of future information used in the imputation via a Bayesian consensus (between posterior distributions) in an optimal way that trades off between the look-ahead-bias and variance.
- We derive a tractable two-step optimization procedure to find the optimal consensus posterior using the Wasserstein barycenter interpolation. The first step finds the best possible distribution to summarize the different posterior models for a given set of weights. The second step optimizes these weights according to the bias-variance criterion that minimizes the variance while explicitly controlling the look-ahead-bias.
2. Related Literature

We briefly review the current imputation methods. Broadly speaking, the various imputation methods can be classified into “single imputation” versus “multiple imputation”. Single imputation refers to the use of a specific number (i.e., a best guess) in place of each missing value, which includes for example, k-nearest neighbor, mean/median imputation, smoothing, interpolation [21] and spline. Matrix completion and matrix factorization [5, 6, 7, 24, 20] are often used to impute panel data, as well as latent factor modelling [10, 2, 33]. Recently, there is a surge in the application of the recurrent neural network for imputation [9, 11, 22, 8], as well as the generative adversarial network [34].

Multiple imputation generalizes the single imputation procedures in that the missing entries are filled-in with multiple guesses instead of one single best guess or estimate, accounting for the uncertainty involved in imputation. Multiple imputation is first developed for non-response in survey sampling [28]. Since then it has been extended to time-series data [17, 16], where a key new element is to preserve longitudinal consistency in imputation. [17] uses smooth basis functions to increase smoothness of imputation. [16] uses gap-filling sequence of imputation for categorical time-series data and produces smooth patterns of transitions. The flexibility and predictive performance of multiple imputation method have been successfully demonstrated in a variety of data sets, such as Industry and Occupation Codes [13], GDP in African nation [17], and concentrations of pollutants in the arctic [18].

The theory of multiple imputation is based on the ignorability assumption, i.e., the data is missing at random. [14] found that earnings data is not missing at random, and hence there is a bias (which they call attrition bias) in estimates of earnings properties. [14] proposed an approach when the random missing pattern depends on observable variables. [19, 26, 31] use logistic regression models to model the missing data distribution.

Notation. For any integers $K \leq L$, we use $[K]$ to denote the set $\{1, \ldots, K\}$ and $[K, L]$ to denote the set $\{K, \ldots, L\}$. We use $\mathcal{M}$ to denote the space of all probability measures supported on $\mathbb{R}^n$. The set of all $n$-dimensional Gaussian measures is denoted by $\mathcal{N}_n$, and we use $\mathcal{S}^n_{++}$ to represent the set of positive-definite matrices of dimension $n \times n$.

3. The Bias-Variance Multiple Imputation Framework

We assume a universe of $n$ assets over $T$ periods. The joint return of these assets at any specific time $t$ is represented by a random vector $Z_t \in \mathbb{R}^n$. The random vector $M_t \in \{0, 1\}^n$ indicates the pattern of the missing values at time $t$, that is, for all $i \in [n]$:

$$\left(M_t\right)_i = \begin{cases} 1 & \text{if the } i\text{-th component of } Z_t \text{ is missing}, \\ 0 & \text{if the } i\text{-th component of } Z_t \text{ is observed.} \end{cases}$$

Coupled with the indicator random variables $M_t$, we define the following projection operator $\mathcal{P}_{M_t}$, that projects the latent variables $Z_t$ to its unobserved components $Y_t$:

$$\mathcal{P}_{M_t}(Z_t) = Y_t.$$

Consequently, the orthogonal projection $\mathcal{P}_{M_t}^\perp$ projects $Z_t$ to the observed components $X_t$

$$\mathcal{P}_{M_t}^\perp(Z_t) = X_t.$$
At the fundamental level, we assume the following generative model

$$\forall t:\begin{cases} Z_t = \theta + \varepsilon_t \\ X_t = \mathcal{P}_{\mathcal{M}_t}(Z_t), \quad Y_t = \mathcal{P}_M(Z_t), \end{cases}$$  

where $\theta, \varepsilon_t \in \mathbb{R}^n$ and $M_t \in \{0,1\}^n$. When the generative model (1) is executed over the $T$ periods, the observed variables $(X_t, M_t)$ are accumulated to produce a data set $\mathcal{D}$ of $n$ rows and $T$ columns. This data set is separated into a training set consisting of $T^{\text{train}}$ periods, and the remaining $T^{\text{test}} = T - T^{\text{train}}$ periods are used as the testing set. We will be primarily concerned with the imputation of the training set.

In this paper, we pursue the Bayesian approach: we assume that the vector $\theta \in \mathbb{R}^n$ is unknown, and is treated as a Bayesian parameter with a prior distribution $\pi_0$; we also treat $(Y_t)_{t \in [T]}$ as Bayesian parameters with appropriate priors. Based on the data set with missing entries $\mathcal{D}$, our strategy relies on computing the distribution of $(Y_t)_{t \in [T]}$ conditional on $\mathcal{D}$, and then generate multiple imputations of $(Y_t)_{t \in [T]}$ by sampling from this distribution. To compute the posterior of $(Y_t)_{t \in [T]}|\mathcal{D}$, we first need to calculate a posterior distribution of the unknown mean vector $\theta$ conditional on $\mathcal{D}$, and then the distribution of $(Y_t)_{t \in [T]}|\mathcal{D}$ is obtained by marginalizing out $\theta$ from the distribution of $(Y_t)_{t \in [T]}|\theta, \mathcal{D}$.

We endeavor in this paper to explore a novel approach to generate multiply-imputed versions of the data set. We make the following assumptions.

**Assumption 3.1.** We observe at least one non-missing value for each row of $\mathcal{D}$ in the training section of $\mathcal{D}$.

**Assumption 3.2.** The missing pattern satisfies the ignorability assumption [23], namely, the probability of $(M_t)_t = 1 \forall t \in [n] \forall t \in [T]$ does not depend on $(Y_t)_{t \in [T]}$ or $\theta$, conditional on $(X_t)_{t \in [T]}$.

**Assumption 3.3** (Bayesian conjugacy). The noise $\varepsilon_t \forall t \in [T]$ in the latent generative model (1) is independently and identically distributed as $\mathcal{N}_n(0, \Omega)$, where $\Omega \in \mathbb{S}_+^n$ is a known covariance matrix. The prior $\pi_0$ is either a non-informative flat prior or Gaussian with mean vector $\mu_0 \in \mathbb{R}^n$ and covariance matrix $\Sigma_0 \in \mathbb{S}_+^n$. The priors on $Y_t$ conditional on $\theta$ are independent across $t$, and the conditional distribution is Gaussian $\mathcal{N}_{\text{dim}(Y)}(\theta Y_t, \Omega Y_t) \forall t \in [T]$, where $\theta Y_t = \mathcal{P}_M(\theta)$, and $\Omega Y_t$ is obtained by applying the projection operator $\mathcal{P}_M$ on $\Omega$.

Using the construction of $\mathcal{P}_M$ in Footnote 1, we have $\Omega Y_t = \mathcal{P}_M \Omega \mathcal{P}_M^T$. Likewise for $\Omega X_t$. The latent generative model (1) implies that the likelihood of $X_t$ conditional on $(\theta, Y_t)$ is Gaussian $\mathcal{N}_{\text{dim}(X)}(\theta X_t + \Omega X_t Y_t^{-1}(Y_t - \theta Y_t), \Omega X_t - \Omega X_t Y_t^{-1} \Omega Y_t X_t)$.

$\theta X_t = \mathcal{P}_{\mathcal{M}_t}(\theta)$, the matrix $\Omega Y_t X_t \in \mathbb{R}^{\text{dim}(Y) \times \text{dim}(X_t)}$ is the covariance matrix between $Y_t$ and $X_t$ induced by $\Omega$, similarly for $\Omega X_t Y_t$.
Figure 2. The Bias-Variance Multiple Imputation (BVMI) framework (contained in dashed box) receives a data set with missing values as an input, and generates multiple data sets as outputs. The posterior generator (G) can exploit time dependency, while the consensus mechanism (C) is bias-variance targeted.

Following our discussion on the look-ahead-bias, we can observe that the distribution of \( \theta|D \) may carry look-ahead-bias originating from the testing portion of \( D \), and this look-ahead-bias will be internalize into the look-ahead-bias of \( (Y_t)_{t \in [T_{train}]}|D \). To mitigate this negative effect, it is crucial to re-calibrate the distribution of \( \theta|D \) to reduce the look-ahead-bias on the posterior distribution of the mean parameter \( \theta \), with the hope that this mitigation will be spilt-over to a similar mitigation of the look-ahead-bias on the imputation of \( (Y_t)_{t \in [T_{train}]} \). Our BVMI framework pictured in Figure 2 is designed to mitigate the look-ahead-bias by implementing there modules:

- (G) The posterior Generator
- (C) The Consensus mechanism
- (S) The multiple imputation Sampler

For the rest of this section, we will elaborate on the generator (G) in Subsection 3.1 and we detail the sampler (S) in Subsection 3.2. The construction of the consensus mechanism (C) is more technical, and thus the full details on (C) will be provided in subsequent sections.

### 3.1. Posterior Generator

We now discuss in more detail how to generate multiple posterior distributions by taking into account the time-dimension of the data set. For convenience we denote \( T_1 = T_{train} \) and \( T_2 = T \). The data set truncated to time \( T_k \), \( k = 1, 2 \) is denoted as \( D_k = \{(X_t, M_t), t \in [T_k]\} \), so that \( D_1 \) coincides with the training data set while \( D_2 \) coincides with the whole data set \( D \). The \( k \)-th posterior for \( \theta \), denoted by \( \pi_k \), is formulated conditional on \( D_k \). The following theorem shows that \( \pi_k \) are Gaussians under Assumptions 3.1 to 3.3. All proofs in this paper are relegated to the supplemental material.

**Theorem 3.4.** Under Assumptions 3.1 to 3.3, the posterior distribution of \( \theta|D_k \) is governed by \( \pi_k \sim N_n(\mu_k, \Sigma_k) \), \( k = 1, 2 \), where if \( \pi_0 \) is non-informative, then

\[
\Sigma_k = \left( \sum_{t \in [T_k]} (P_{M_t}^{-1})(\Omega_{X_t}^{-1})^{-1} \right)^{-1},
\]

\[
\mu_k = \Sigma_k \left( \sum_{t \in [T_k]} (P_{M_t}^{-1})(\Omega_{X_t}^{-1})(P_{M_t}^{-1})(X_t) \right).
\]
and if \( \pi_0 \) is \( N_n(\mu_0, \Sigma_0) \), then

\[
\Sigma_k = \left( \Sigma_0^{-1} + \sum_{t \in [T_k]} (P_{M_t}^\perp)^{-1}(\Omega_{X_t}^{-1}) \right)^{-1}
\]

\[
\mu_k = \Sigma_k \left( \mu_0 + \sum_{t \in [T_k]} (P_{M_t}^\perp)^{-1}(\Omega_{X_t}^{-1})(P_{M_t}^\perp)^{-1}(X_t) \right),
\]

(3)

where \((P_{M_t}^\perp)^{-1}\) is the inverse map of \( P_{M_t}^\perp \) obtained by filling entries with zeroes.

The product of the posterior generator module \((G)\) is two elementary posteriors \( \{\pi_k\}_{k \in [2]} \), where \( \pi_k \) is the posterior of \( \theta|D_k \). The definition of the set \( D_1 \) and \( D_2 \) implies that the posterior \( \pi_2 \) carries look-ahead-bias relative to the posterior \( \pi_1 \), meanwhile also has less variance. The two posteriors are transmitted to the consensus mechanism \((C)\) to form an aggregated posterior \( \pi^* \) that strikes a balance between the look-ahead-bias and variance. The aggregated posterior \( \pi^* \) is injected to the sampler module \((S)\), which we now study.

3.2. Multiple Imputation Sampler

A natural strategy to impute the missing values depends on recovering the (joint) posterior distribution of \((Y_t)_{t \in [T]}\) conditional on the observed data \( D \). Fortunately, if the noise \( \epsilon_t \) are mutually independent and the priors on \( Y_t \) are mutually independent given \( \theta \), it suffices to consider the posterior distribution of \( Y_t \) for any given \( t \).

**Proposition 3.5.** Under Assumption \( \beta \beta \), the distribution of \( Y_t \) conditional on \( \theta \) and \((X_t, M_t)_{t \in [T]}\) is independent across \( t \), and

\[
Y_t | \theta, (X_t, M_t)_{t \in [T]} \sim N_{\dim(Y_t)}(\tilde{\theta}_t, \tilde{\Omega}_t),
\]

(4)

where \( \tilde{\theta}_t = \theta_{Y_t} + \Omega_{Y_tX_t} \Omega_{X_t}^{-1}(X_t - \theta_{X_t}) \) and \( \tilde{\Omega}_t = \Omega_{Y_t} - \Omega_{Y_tX_t} \Omega_{X_t}^{-1} \Omega_{X_tY_t} \).

Proposition 3.5 indicates that the distribution of \( Y_t|D \) coincide with the distribution of the random vector \( \xi_t \in \mathbb{R}^{\dim(Y_t)} \) dictated by

\[
\xi_t = A_t \theta + b_t + \eta_t, \theta \sim \pi^*, \eta_t \sim N_{\dim(Y_t)}(0, S_t), \theta \perp \eta_t,
\]

(5)

for some appropriate parameters \( A_t \), \( b_t \) and \( S_t \) that are designed to match the mean vector and the covariance matrix of the Gaussian distribution in [4].

The variability in the sampling of \( \xi_t \) using [5] comes from two sources: that of sampling \( \theta \) from \( \pi^* \) and that of sampling \( \eta_t \) from \( N_{\dim(Y_t)}(0, S_t) \). As we have seen in Theorem 3.4, the posterior covariance of \( \theta \) is inverse-proportional to the time-dimension of the data set, and therefore the variability due to \( \pi^* \) is likely to be overwhelmed by the variability due to \( \eta_t \) unless \( T \) is small. To fully exhibit the potential of the aggregated posterior \( \pi^* \) for look-ahead-bias and variance trade-off, especially when validating our proposed method in numerical experiments, we choose to eliminate the idiosyncratic noise \( \eta_t \), namely, the missing values are imputed by its conditional expectation given \( \theta \), which is depicted in Algorithm [4].

4. Bias-Variance Targeted Consensus Mechanism

We devote this section to provide the general technical details on the construction of the consensus mechanism module \((C)\) in our BVMI framework. Conceptually, the consensus mechanism module receives two

---

3This inverse map can be defined as \( (P_{M_t}^\perp)^{-1}(\Omega_{X_t}^{-1}) = \arg\min \{\|\tilde{\Omega}\|_0 : \tilde{\Omega} \in \mathbb{R}^{n \times n}, P_{M_t}^\perp(\tilde{\Omega}) = \Omega_{X_t}^{-1}, \} \), where \( \|\cdot\|_0 \) is a sparsity inducing norm. Similarly for \( (P_{M_t}^\perp)^{-1}(X_t) \).
Algorithm 1 Conditional Expectation Bayesian Imputation

**Input:** aggregated posterior $\pi^*$, covariance matrix $\Omega$, data set $\mathcal{D} = \{(X_t, M_t) : t \in [T]\}$,

Sample $\theta$ from $\pi^*$

for $t = 1, \ldots, T$

Compute $A_t$, $b_t$ for model (5) using $(X_t, M_t, \Omega)$

Impute $Y_t$ by $A_t\theta + b_t$

end for

**Output:** An imputed data set

Algorithm 1: Conditional Expectation Bayesian Imputation

---

Posteriors $\pi_1, \pi_2$ from the posterior generator module (G), then synthesizes a unique posterior $\pi^*$ and transmits $\pi^*$ to the sampler module (S). We provide a formal definition of a consensus mechanism. First, we define the Wasserstein distance.

**Definition 4.1** (Wasserstein distance). For $p \geq 1$ the $p$-Wasserstein distance for two probability measures $\pi_1, \pi_2 \in \mathcal{M}$ is

$$W_p(\pi_1, \pi_2) = \left(\inf_{\pi \in \Pi(\pi_1, \pi_2)} \int_{\mathbb{R}^{2n}} \|x - y\|_2^p d\pi(x, y)\right)^{1/p}.$$

where $\Pi(\pi_1, \pi_2)$ is the set of all probability measures on $\mathbb{R}^{2n}$ that have marginals $\pi_1$ and $\pi_2$.

We focus on the 2-Wasserstein distance due to its analytical tractability. Utilizing this distance, we define the Wasserstein consensus as follows.

**Definition 4.2** (Consensus mechanism). The Wasserstein consensus mechanism $C_\lambda : \mathcal{M}^2 \to \mathcal{M}$ parametrized by $\lambda$ outputs a unifying posterior $\pi^*$ as

$$C_\lambda(\{\pi_k\}) = \arg \min_{\pi \in \mathcal{M}} \lambda_1 W_2^2(\pi, \pi_1) + \lambda_2 W_2^2(\pi, \pi_2).$$

The space of all possible consensus mechanism induced by the simplex $\Delta = \{\lambda \in \mathbb{R}^2_+ : \lambda_1 + \lambda_2 = 1\}$ is represented by $C = \{C_\lambda : \lambda \in \Delta\}$. Wasserstein distance enjoys many sound properties, both in theory and in practice.

Problem (6) computes the probability measure that minimizes the sum of the distances to a set of target measures. The optimal solution is usually called the Wasserstein barycenter, or McCann’s interpolation in the case of two target measures. The parameter vector $\lambda$ can be thought of as the vector of weights assigning to the set of posteriors $\{\pi_k\}$. Problem (6) is infinite-dimensional as the decision variable $\pi$ ranges over the space of all probability distributions supported on $\mathbb{R}^n$. Fortunately, when the posteriors $\pi_k$ are Gaussian, it is well documented that $C_\lambda$ can be characterized analytically.

**Proposition 4.3.** Suppose that $\pi_k \sim \mathcal{N}(\mu_k, \Sigma_k)$ with $\Sigma_k > 0$ for $k = 1, 2$. For any $\lambda \in \Delta$, we have $C_\lambda(\{\pi_k\}) = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$ with

$$\hat{\mu} = \lambda_1 \mu_1 + \lambda_2 \mu_2 \in \mathbb{R}^n,$$

$$\hat{\Sigma} = (\lambda_1 I_n + \lambda_2 \Phi) \Sigma_1 (\lambda_1 I_n + \lambda_2 \Phi) \in \mathbb{S}_{++}^n,$$

where $\Phi = \Sigma_2^{1/2} \left(\Sigma_1^{1/2} \Sigma_2^{1/2} \Sigma_1^{1/2}\right)^{-1/2} \Sigma_2^{1/2}$.

Next we identify the optimal consensus mechanism $C_{\lambda^*} \in C$ according to the notion of bias-variance trade-off, which is ubiquitously used in (Bayesian) statistics. The first step towards this goal is to quantify the variance of the aggregated posterior. Let $\text{Var}_{C_{\lambda^*}}(\{\pi_k\}) \in \mathbb{R}^n$ be the vector of variances of $\theta$ under the posterior distribution $C_{\lambda^*}(\{\pi_k\})$. We consider the following minimal variance criterion which is computed based on taking the sum of the individual variances.
**Definition 4.4** (Minimal variance posterior). For a fixed input \( \{\pi_k\} \), the aggregated posterior \( \pi^* \) obtained by \( \pi^* = C_{\lambda^*}(\{\pi_k\}) \) for an \( C_{\lambda^*} \in \mathbb{C} \) has minimal variance if
\[
1^T \text{Var}_{\pi^*}(\theta) \leq 1^T \text{Var}_{C_{\lambda^*}(\{\pi_k\})}(\theta) \quad \forall \mathbb{C} \in \mathbb{C},
\]
where \( 1 \) is the vector of ones.

The second steps entails quantifying the bias of an aggregated posterior \( \pi^* \). Ideally, this bias quantity is measured by
\[
\|E_{\pi^*}(\theta) - \theta^{\text{true}}\|_2,
\]
where \( \theta^{\text{true}} \) is the true mean of the returns vector. Unfortunately, the value \( \theta^{\text{true}} \) is not available in practice, and hence quantifying the true bias of an aggregated posterior is impossible. As an alternative, we resort to Theorem 4.7.

**Theorem 4.7.** Suppose that \( C_{\lambda} \) for a fixed weight parameter \( \lambda \), is optimal for \( \{\pi_k\} \).

**Definition 4.5** ((\( \mu, \delta \))-bias tolerable posterior). Fix an anchored mean \( \mu \in \mathbb{R}^n \) and a tolerance \( \delta \in \mathbb{R}_+ \). A posterior \( \pi^* \) is said to be \( (\mu, \delta \) -bias tolerable if
\[
\|E_{\pi^*}(\theta) - \mu\|_2 \leq \delta.
\]

The tolerance parameter \( \delta \) indicates how much of bias is accepted in the aggregated posterior \( \pi^* \). Because the ultimate goal of the consensus mechanism is to mitigate the look-ahead-bias, it is imperative to choose the anchored mean \( \mu \) in an appropriate manner. As the posterior \( \pi_1 \) obtained by conditioning on the training set \( D_1 = \{(X_t, M_t) : t \in [T_{\text{train}}]\} \) carries the least amount of future information, it is reasonable to consider the discrepancy between the mean of the aggregated posterior \( \pi^* \) and that of \( \pi_1 \) as a proxy for the true bias. As a result we have the following notion of an optimal consensus mechanism.

**Definition 4.6** (Optimal consensus mechanism). For a given input \( \{\pi_k\} \). The consensus mechanism \( C_{\lambda^*} \in \mathbb{C} \) is optimal for \( \{\pi_k\} \) if \( \lambda^* \) solves
\[
\min_{\lambda \in \Delta} \|E_{C_{\lambda^*}(\{\pi_k\})}(\theta) - E_{\pi_1}(\theta)\|_2 \leq \delta.
\]

Intuitively, the optimal consensus mechanism for the set of elementary posteriors \( \{\pi_k\} \) is chosen so as to generate an aggregated posterior \( \pi^* \) that has low variance and acceptable level of look-ahead-bias. Notice that problem 8 shares significant resemblance with the Markowitz’s problem: the objective function of 8 minimizes a variance proxy, while the constraint of 8 involves the expectation parameters. It is important to notice the sharp distinction between our consensus optimization problem 8 and the existing literature. While the existing literature typically focuses on studying the barycenter for a fixed weight parameter \( \lambda \), problem 8 searches for the best value of \( \lambda \) that meets our optimality criteria.

Proposition 4.3 provides us with a tractable reformulation of problem 8, which is the main result of this section.

**Theorem 4.7.** Suppose that \( \pi_k \sim \mathcal{N}(\mu_k, \Sigma_k) \) with \( \Sigma_k > 0 \) for \( k = 1, 2 \). Let \( \lambda^* \) be the optimal solution in the variable \( \lambda \) of the following quadratic programming
\[
\min_{\lambda \in \Delta} \text{Tr}(\Sigma_1)\lambda_1^2 + \text{Tr}(\Sigma_2)\lambda_2^2 + 2\text{Tr}(\Sigma_1\Phi)\lambda_1\lambda_2
\]
\[
\text{s.t.} \quad \lambda_2\|\mu_1 - \mu_2\|_2 \leq \delta,
\]
where \( \Phi = \Sigma_2^{-1/2} \left( \Sigma_2^{-1/2} \Sigma_1^{-1/2} \right)^{-1/2} \Sigma_2^{1/2} \). Then \( C_{\lambda^*} \) is the optimal consensus mechanism in \( \mathbb{C} \).
5. Numerical Experiment

In this section we design and carry out numerical experiments to validate the look-ahead-bias and variance trade-off that is captured by our proposed BVMI framework. All optimization problems are modeled in Python and solved by the solver GUROBI on an Intel Core i5 CPU (1.40 GHz) computer.

The panel of asset returns data (either synthetic or real) is separated into three portions: training, testing, and an additional out-of-sample testing periods, each having a length \( T_{\text{train}}, T_{\text{test}}, \text{and } T_{\text{oos-test}} \) respectively. In both synthetic and real data experiments we will manually create missingess by applying masks to entries in the panel following certain missing mechanisms. We only apply the masking procedure to the training period, while we preserve the completeness of testing and out-of-sample testing periods for evaluation purposes.

Following the BVMI framework, we feed the data set with missing values to the posterior generator module, as we have detailed in Section \[3.1\]. Recall that we have a total of two posterior distributions, which are fitted using the training period and training plus testing periods respectively. We assume a non-informative flat prior \( \pi_0 \) on the mean parameter \( \theta \). The covariance matrix \( \Omega \) is taken as the ground truth in the case of synthetic data experiment, or fitted from the data in the case of real data experiment. The generated multiple posterior distributions are then fed to the consensus mechanism module to output an optimal consensus posterior, the details of which we have provided in Sections \[4\]. As we have remarked before, the tolerance parameter \( \delta \) in \([5]\) (or equivalently in \([6]\)) indicates how much of look-ahead-bias is accepted in the aggregated posterior \( \pi^* \). We experiment with the \( \delta \) parameter selected to be 10 values equi-spaced between 0 and \( \delta_{\text{max}} \), where \( \delta = 0 \) gives us the aggregated posterior \( \pi^* \) matching the posterior \( \pi_1 \), and \( \delta = \delta_{\text{max}} \) gives us the aggregated posterior \( \pi^* \) matching the posterior \( \pi_2 \). We can compute \( \delta_{\text{max}} \) by \( \|\mu_1 - \mu_2\|_2 \). Corresponding to each selected \( \delta \) parameter, the aggregated posterior \( \pi^* \) is then output to the multiple imputation sampler module, where we follow Algorithm \[2\] to create multiply-imputed data sets.

The downstream task to be performed on the imputed data set is that of portfolio optimization. We consider a greedy strategy to construct the portfolio weights, that is, we calculate the \( \langle w, Z_t \rangle \) matching the posterior \( \pi^* \), and \( \langle w, Z_t \rangle \) matching the posterior \( \pi_1 \) and \( \pi_2 \). Corresponding to each selected \( \delta \) parameter, the aggregated posterior \( \pi^* \) is then output to the multiple imputation sampler module, where we follow Algorithm \[2\] to create multiply-imputed data sets.

The difference in averaged daily portfolio return between the two periods, or \( \text{regret} \), is defined as \( \Delta R = R_{\text{test}} - R_{\text{oos-test}} \).
Note that we consider one-sided bias in the above formulation. We call the first term above as ECBias$^2$, and the second term above as ECVar.

In the following, we make further specifications of the experiments depending on whether we work with synthetic or real data. We present the (qualitative) trend of ECMSE, ECBias$^2$ and ECVar as $\delta$ varies, under three different missing mechanisms which include missing at random, a dependent block missing structure, and when the missing pattern depends on realizations of returns. The results show that our proposed BVMI framework has the potential to outperform naive posteriors in terms of ECMSE.

5.1. Experiments with Synthetic Data

We consider number of stocks $n = 100$, and horizons $T^{\text{train}} = T^{\text{test}} = 100$, $T^{\text{oos-test}} = 1000$. We generate the returns vector from a factor model, namely, the vectors $Z_t, t = 1, \ldots, T^{\text{train}} + T^{\text{test}} + T^{\text{oos-test}}$ are generated i.i.d. from $N_n(\theta, \Omega)$, where for $i = 1, \ldots, n, \theta_i = 0.2\beta_i + \alpha_i$, $\beta_i = 1$, $\alpha_i$ is equi-spaced between $-0.3$ and $0$. $\beta$ and $\Omega = \beta\beta^\top + I_n$.

We consider three types of missing mechanisms: missing completely at random - we mask each entry in the training period as missing independently at random with a fixed probability 70%; block missing - we mask the first 60% of the training period as missing; missing by value - we mask an entry in the training period as missing if its absolute value is greater than 0.25.

For each missing mechanism, we repeat 100 independent experiments, where in each experiment the sampled panel is multiply-imputed 100 times, and we carry out the downstream task on each imputed data set. The trends of ECMSE, ECBias$^2$ and ECVar with respect to the $\delta$ parameter are summarized in Figure 3. We see that in all three cases of missing mechanisms, the bias component ECBias$^2$ is increasing as $\delta/\delta_{\text{max}}$ transitions from 0 to 1, while the variance component ECVar is decreasing. For the first two cases, the sum of the two components ECMSE achieves a minimum for some non-trivial selection of $\delta$, showing that our method outperforms the naive posteriors (i.e., the posteriors corresponding to $\delta = 0$ or $\delta = \delta_{\text{max}}$).

5.2. Experiments with Real Data

We choose the data as the $n = 49$ industry portfolio returns downloaded from Kenneth French’s website in a duration from Jan. 3rd, 2017 to Dec. 31st, 2018. We consider $T^{\text{train}} = 200$, $T^{\text{test}} = 100$ and $T^{\text{oos-test}} = 100$. The covariance matrix $\Omega$ in Assumption 3.3 is fitted using the complete panel (a total of 502 days). As before, we consider three types of missing mechanisms. However, we use the data set differently in each case.

- **Missing completely at random** - we mask each entry in the training period as missing independently at random with a fixed probability 70%. The data is fixed as the first $T^{\text{train}} + T^{\text{test}} + T^{\text{oos-test}} = 400$ dates from the original data set. We repeat 50 independent experiments, where in each experiment we first generate the missing masks and then multiply-impute 100 times.

- **Block missing** - we mask the first 65% of the training period as missing. We conduct 50 independent experiments. In the $k$-th experiment, we use the $T^{\text{train}} + T^{\text{test}} + T^{\text{oos-test}} = 400$ days of data starting from the $k$-th date in the original data set, and we multiply-impute 100 times.

- **Missing by value** - we mask an entry in the training period as missing if its absolute value is greater than 0.1. The rest of the procedure is the same as the previous case.

The trends of ECMSE, ECBias$^2$ and ECVar with respect to the $\delta$ parameter are summarized in Figure 4. We see that in all three cases of missing mechanisms, the bias component ECBias$^2$ is increasing as $\delta/\delta_{\text{max}}$ transitions from 0 to 1, while the variance component ECVar is decreasing. The sum of the two components ECMSE achieves a minimum for some non-trivial selection of $\delta$ for the last two cases, showing that our method has the potential to outperform the naive posteriors (i.e., the posteriors corresponding to $\delta = 0$ or $\delta = \delta_{\text{max}}$).

---

Retrieved on Jan. 24th, 2021 from [https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html](https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html)
Figure 3. Experimental results on synthetic data set

**Acknowledgments.** We gratefully acknowledge the financial support and feedback of MSCI and its data science team.
Figure 4. Experimental results on real data set

References

[1] M. Agueh and G. Carlier, *Barycenters in the wasserstein space*, SIAM Journal on Mathematical Analysis, 43 (2011), pp. 904–924.
[2] J. Bai and S. Ng, Matrix completion, counterfactuals, and factor analysis of missing data, arXiv:1910.06677, (2020).
[3] R. W. Banz and W. J. Breen, Sample-dependent results using accounting and market data: Some evidence, The Journal of Finance, 41 (1986), pp. 779–793.
[4] C. Breunig, M. Kummer, J. Oehnenus, and S. Viete, Information technology outsourcing and firm productivity: eliminating bias from selective missingness in the dependent variable, The Econometrics Journal, 23 (2019), pp. 88–114.
[5] J.-F. Cai, E. J. Candès, and Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM Journal on Optimization, 20 (2010), pp. 1956–1982.
[6] E. J. Candès and B. Recht, Exact matrix completion via convex optimization, Foundations of Computational Mathematics, 9 (2009), p. 717.
[7] E. J. Candès and T. Tao, The power of convex relaxation: Near-optimal matrix completion, IEEE Transactions on Information Theory, 56 (2010), pp. 2053–2080.
[8] W. Cao, D. Wang, J. Li, H. Zhou, Y. Li, and L. Li, Britx: Bidirectional recurrent imputation for time series, in Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, p. 6776–6786.
[9] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, Recurrent neural networks for multivariate time series with missing values, Nature Scientific Reports, 8 (2018), p. 6085.
[10] Y. Chen, J. Fan, C. Ma, and Y. Yan, Inference and uncertainty quantification for noisy matrix completion, Proceedings of the National Academy of Sciences, 116 (2019), pp. 22931–22937.
[11] E. Choi, M. T. Bahadori, A. Schultz, W. F. Stewart, and J. Sun, Doctor ai: Predicting clinical events via recurrent neural networks, vol. 56 of Proceedings of Machine Learning Research, 2016, pp. 301–318.
[12] J. Cho, C. Liu, and J. Woo, An efficient approach for removing look-ahead bias in the least square monte carlo algorithm: Leave-one-out, arXiv:1810.02071, (2018).
[13] C. C. Clogg, D. B. Rubin, N. Schenker, B. Schultz, and L. Weidman, Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression, Journal of the American Statistical Association, 86 (1991), pp. 68–78.
[14] P. Easton, M. Kapons, P. Kelly, and A. Neuhaefer, Attrition bias and inferences regarding earnings properties; evidence from compustat data, SSRN, (2020).
[15] H. Guo, H. Kassa, and M. F. Ferguson, On the relation between EGARCH idiosyncratic volatility and expected stock returns, Journal of Financial and Quantitative Analysis, 49 (2014), p. 271–296.
[16] B. Halpin, Multiple imputation for categorical time series, The Stata Journal, 16 (2016), pp. 590–612.
[17] J. Honaker and G. King, What to do about missing values in time series cross-section data, American Journal of Political Science, 54 (2010), pp. 561–581.
[18] P. K. Hopke, C. Liu, and D. B. Rubin, Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the arctic, Biometrics, 57 (2001), pp. 22–33.
[19] J. G. Ibrahim, S. R. Lipsitz, and M.-H. Chen, Missing covariates in generalized linear models when the missing data mechanism is non-ignorable, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61 (1999), pp. 173–190.
[20] R. H. Keshavan, A. Montanari, and S. Oh, Matrix completion from a few entries, IEEE Transactions on Information Theory, 56 (2010), pp. 2980–2989.
[21] D. M. Kreindler and C. J. Lumsden, The effects of the irregular sample and missing data in time series analysis, Nonlinear Dynamics Psychol Life Sci, 10 (2006), pp. 187–214.
[22] Z. C. Lipton, D. Kale, and R. Wettzel, Directly modeling missing data in sequences with rnns: Improved classification of clinical time series, vol. 56 of Proceedings of Machine Learning Research, 2016, pp. 253–270.
[23] R. Little and D. Rubin, Statistical analysis with missing data, Wiley, 2002.
[24] R. Mazumder, T. Hastie, and R. Tibshirani, Spectral regularization algorithms for learning large incomplete matrices, Journal of Machine Learning Research, 11 (2010), pp. 2287–2322.
[25] R. J. McCann, A convexity principle for interacting gases, Advances in Mathematics, 128 (1997), pp. 153–179.
[26] K. Morikawa, J. K. Kim, and Y. Kano, Semiparametric maximum likelihood estimation with data missing not at random, 45 (2017), pp. 393–409.
[27] O. Pele and M. Werman, Fast and robust earth mover’s distances, in International Conference on Computer Vision, vol. 9, 2009, pp. 460–467.
[28] D. B. Rubin, Multiple Imputation for Nonresponse in Surveys, Wiley, 1987.
[29] Y. Rubner, L. Guibas, and C. Tomasi, The earth mover’s distance, multi-dimensional scaling, and color-based image retrieval, in Proceedings of the ARPA Image Understanding Workshop, 1997, pp. 661–668.
[30] M. J. Schervish, Theory of Statistics, Springer, 1995.
[31] A. Sportisse, C. Boyer, and J. Josse, Imputation and low-rank estimation with missing not at random data, Statistics and Computing, 30 (2020), pp. 1629–1643.
[32] C. Villani, Optimal Transport: Old and New, Springer Science & Business Media, 2008.
[33] R. Xiong and M. Pelger, Large dimensional latent factor modeling with missing observations and applications to causal inference, SSRN, (2020).
Proof of Theorem 3.4. Let \( f(D_k|\theta) \) be the conditional density representing the likelihood of observing \( D_k \) given the parameter \( \theta \). Under Assumption 3.2, we have
\[
 f(D_k|\theta) = f_X((X_t)_{t \in [T_k]}|\theta) \times f_M((M_t)_{t \in [T_k]}|X_t), \theta) = f_X((X_t)_{t \in [T_k]}|\theta) \times \tilde{f}_M((M_t)_{t \in [T_k]}|X_t)
\]
for some appropriate (conditional) likelihood functions \( f_X, f_M \) and \( \tilde{f}_M \). Now the Bayes theorem \([30]\) Theorem 1.31] asserts that
\[
 \frac{d\pi_k}{d\pi_0}(\theta|D_k) = \frac{f(D_k|\theta)}{\int_{\mathbb{R}^n} f(D_k|\theta) \pi_0(d\theta)} = \frac{f_X((X_t)_{t \in [T_k]}|\theta)}{\int_{\mathbb{R}^n} f_X((X_t)_{t \in [T_k]}|\theta) \pi_0(d\theta)}
\]
where \( d\pi_k/d\pi_0 \) represents the Radon-Nikodym derivative of \( \pi_k \) with respect to \( \pi_0 \). Under Assumption 3.3 \( X_t|\theta \) is Gaussian with mean \( \theta_X \) and covariance matrix \( \Omega_X \). Thus the posterior \( \pi_k \) satisfies
\[
 \frac{d\pi_k}{d\pi_0}(\theta|D_k) = \prod_{t \in [T_k]} \mathcal{N}_{\text{dim}(X_t)}(X_t|\theta_X,\Omega_X) \propto \exp\left(-\frac{1}{2}(\theta_0 - \mu_0)^\top \Sigma_0^{-1}(\theta_0 - \mu_0)\right) \prod_{t \in [T_k]} \exp\left(-\frac{1}{2}(\theta_X - X_t)^\top \Omega_X^{-1}(\theta_X - X_t)\right).
\]
By expanding the exponential terms and completing the squares, we see that \( \pi_k \sim \mathcal{N}_n(\mu_k, \Sigma_k) \) with
\[
 \Sigma_k = \left(\Sigma_0^{-1} + \sum_{t \in [T_k]} (P_{M_t}^\perp)^{-1}(\Omega_X^{-1})\right)^{-1}, \quad \mu_k = \Sigma_k \left(\Sigma_0^{-1}\mu_0 + \sum_{t \in [T_k]} (P_{M_t}^\perp)^{-1}(\Omega_X^{-1})(P_{M_t}^\perp)^{-1}(X_t)\right),
\]
Since \( \Omega_X^{-1} \) is positive definite, we have that \( (P_{M_t}^\perp)^{-1}(\Omega_X^{-1}) \) is positive semidefinite. The matrix inversion that defines \( \Sigma_k \) is thus well-defined thanks to the positive definiteness of \( \Sigma_0 \) in Assumption 3.3.

Similarly, if \( \theta \) has a flat uninformative prior, then \( \pi_k \sim \mathcal{N}_n(\mu_k, \Sigma_k) \) where
\[
 \Sigma_k = \left(\sum_{t \in [T_k]} (P_{M_t}^\perp)^{-1}(\Omega_X^{-1})\right)^{-1}, \quad \mu_k = \Sigma_k \left(\sum_{t \in [T_k]} (P_{M_t}^\perp)^{-1}(\Omega_X^{-1})(P_{M_t}^\perp)^{-1}(X_t)\right).
\]
We argue that the matrix inversion that defines \( \Sigma_k \) is well-defined under Assumption 3.1. Let \( \xi \) by any non-zero vector of dimension \( n \). From Assumption 3.1, there exists \( t \in [T_k] \), such that \( \xi_{X_t} = P_{M_t}^\perp(\xi) \) is non-zero. Since \( \Omega_X \) is positive definite, we have \( \xi_{X_t}^\top \Omega_X^{-1} \xi_{X_t} > 0 \). Therefore
\[
 \xi^\top ((P_{M_t}^\perp)^{-1}(\Omega_X^{-1})) \xi > 0.
\]
Thus
\[
 \xi^\top \left(\sum_{t \in [T_k]} (P_{M_t}^\perp)^{-1}(\Omega_X^{-1})\right) \xi > 0.
\]
Since \( \xi \) is arbitrary, we have \( \sum_{t \in [T_k]} (P_{M_t}^\perp)^{-1}(\Omega_X^{-1}) \) is positive definite. Thus \( \Sigma_k \) is positive definite. □
Proof of Proposition 3.5. Under Assumption 3.3, conditional on $\theta$, the joint distribution of $X_t, Y_t$ (i.e. $Z_t$) is Gaussian $\mathcal{N}_n(\theta, \Omega)$ and is independent across $t$. The distribution of $Y_t$ conditional on $\theta$ and $(X_t, M_t)$ is thus a conditional Gaussian given by

$$Y_t|\theta, X_t, M_t = Y_t|\theta, (X_t, M_t)_{t \in [T]} \sim \mathcal{N}_{\text{dim}(Y_t)}(\theta Y_t + \Omega_{Y_t} X_t \Omega_{X_t}^{-1} (X_t - \theta X_t), \Omega_{Y_t} - \Omega_{Y_t} X_t \Omega_{X_t}^{-1} \Omega_{X_t} Y_t).$$

□

Proof of Proposition 4.3. From [1, Section 6.2] the Wasserstein barycenter for two target measures is the well-known McCann’s interpolation [25]. If the two target measures are Gaussians, then [25, Example 1.7] asserts the barycenter to be Gaussian with the desired mean and covariance matrix. □

Proof of Theorem 4.7. From Proposition 4.3, it is straightforward to compute

$$\mathbb{E}^\top \text{Var}_{\mathcal{C}_1((\pi_k))}[\theta] = \text{Tr}(\hat{\Sigma}) = \text{Tr}((\lambda_1 I_n + \lambda_2 \Phi) \Sigma_1 (\lambda_1 I_n + \lambda_2 \Phi)) = \lambda_1^2 \text{Tr}(\Sigma_1) + \lambda_1 \lambda_2 (\text{Tr}(\Sigma_1 \Phi) + \text{Tr}(\Phi \Sigma_1)) + \lambda_2^2 \text{Tr}(\Phi \Sigma_1 \Phi) = \lambda_1^2 \text{Tr}(\Sigma_1) + 2 \lambda_1 \lambda_2 \text{Tr}(\Sigma_1 \Phi) + \lambda_2^2 \text{Tr}(\Sigma_2),$$

and

$$\|\mathbb{E}_{\mathcal{C}_1((\pi_k))}[\theta] - \mathbb{E}_{\pi_1}[\theta]\|_2 = \|\hat{\mu} - \mu_1\|_2 = \|\lambda_1 - 1\| \mu_1 + \lambda_2 \mu_2\|_2 = \lambda_2 \|\mu_1 - \mu_2\|_2.$$}

Substitute the above expressions for problem (8) gives us the desired result. □