Cost-Efficient VBI-Based Multiuser Detection for Uplink Grant-Free MIMO-NOMA

Boran Yang*, Xiaoxu Zhang*, Li Hao*, George K. Karagiannidis†, Xin Quan*, and Octavia A. Dobre‡
*Southwest Jiaotong University, Chengdu, China
†Aristotle University of Thessaloniki, Thessaloniki, Greece and Lebanese American University, Beirut, Lebanon
‡Memorial University, St. John's, Canada
Email: bryang@my.swjtu.edu.cn

2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring) | 979-8-3503-8741-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/VTC2024-SPRING62846.2024.10683341

***Abstract*—Grant-free non-orthogonal multiple access (GF-NOMA) based on multiple-input multiple-output (MIMO) has attracted much attention as a promising technique to support massive connectivity and bursty data transmission in massive machine-type communication. In this paper, we propose two compressed sensing based multiuser detection (MUD) algorithms for the MIMO-enabled GF-NOMA system. First, the spatially enhanced variational Bayesian inference (SE-VBI) algorithm is developed for MUD by exploiting the Gaussian mixture prior and diversity combining technique. Then, by applying the covariance-free (CoFe) strategy to the SE-VBI framework to estimate the diagonal elements of the posterior covariance, we propose a low-complexity MUD method named SE-CoFe-VBI. In particular, the proposed algorithms integrate the multivariate nature of the transmitted signal, i.e., discreteness, sparsity, and spatial correlation. Simulation results show that the proposed algorithms offer improved detection performance over the state-of-the-art spatially enhanced sparse Bayesian learning method.**

***Index Terms*—Grant-free non-orthogonal multiple access, multiple-input multiple-output, Gaussian mixture prior, variational Bayesian inference, covariance-free.**

I. Introduction

With the generational evolution of wireless communication technologies, massive machine-type communication (mMTC) has emerged as a powerful backbone for B5G cellular Internet-of-Things (IoT) [1]. Unfortunately, the current request-grant based orthogonal multiple access protocols cannot meet the requirements of massive connectivity and sporadic service traffic in mMTC [2]. To address this challenge, the grant-free non-orthogonal multiple access (GF-NOMA) solutions have gained prominence as a preferred alternative to mitigate signaling cost and round-trip delay [3]. In GF-NOMA, active users can transmit data to the base station (BS) arbitrarily without any restrictions on resource allocation and scheduling priority. However, how to perform effective multiuser detection (MUD) is the primary task of GF-NOMA. Fortunately, the inherent bursty nature of user traffic allows the use of compressed sensing (CS) techniques to solve the sparse detection problem [4]–[8].

Bayesian-based CS algorithms, such as approximate message passing (AMP) [4] and sparse Bayesian learning (SBL) [5], have been proposed for MUD by using sparsity-promoting prior distributions. However, the works in [4] and [5] failed to explore the impact of degrees of spatial freedom on MUD performance. Motivated by the above observations, the authors in [6] investigated the MUD problem in the multiple-input multiple-output (MIMO) case and proposed a parallel AMP algorithm for multiple measurement vector formulation. The block SBL (B-SBL) [7] algorithm, which assigns a structured prior distribution to the block sparse signal, was developed to perform MUD. On the other hand, our previous work [8] proposed a spatially enhanced sparse Bayesian learning (SE-SBL) algorithm to improve the posterior distribution of the transmitted signal by incorporating the spatial diversity property of multi-antenna reception. However, the aforementioned studies [4]–[8] neglected the potential formats of the transmitted symbols, i.e., zero paradigm for inactive users and finite alphabet for active users.

In this paper, we formulate the MUD for MIMO-enabled GF-NOMA system as a block sparse detection problem and propose two promising Bayesian CS algorithms. First, the spatially enhanced variational Bayesian inference (SE-VBI) method is developed to enable MUD. This algorithm is an improved version of SE-SBL by integrating the discreteness and sparsity of the transmitted symbols for both inactive and active users. Second, we develop a low-cost MUD method by applying the covariance-free (CoFe) technique [9] to the SE-VBI framework, i.e., SE-CoFe-VBI. This algorithm uses the diagonal estimation principle and fundamental matrix theory to approximate the posterior distribution of the transmitted signal. Finally, numerical results show that SE-VBI and SE-CoFe-VBI exhibit similar performance and significantly outperform the state-of-the-art SE-SBL algorithm.

The rest of the paper is organized as follows. Section II introduces the system model for grant-free MIMO-NOMA. Sections III and IV present the SE-VBI and SE-CoFe-VBI algorithms for MUD, respectively. Section V describes the experimental results, and Section VI concludes the paper.

II. System Model

We consider a MIMO-enabled GF-NOMA system that includes a BS with M antennas and K single-antenna users, as shown in Fig. 1(a). It is assumed that only a few active users

This work was partially funded by the National Science Foundation of China Project (No. 62020106001), and in part by the Sichuan Science and Technology Program (No. 2022NSFSC0910).

979-8-3503-8741-4/24/$31.00 ©2024 IEEE

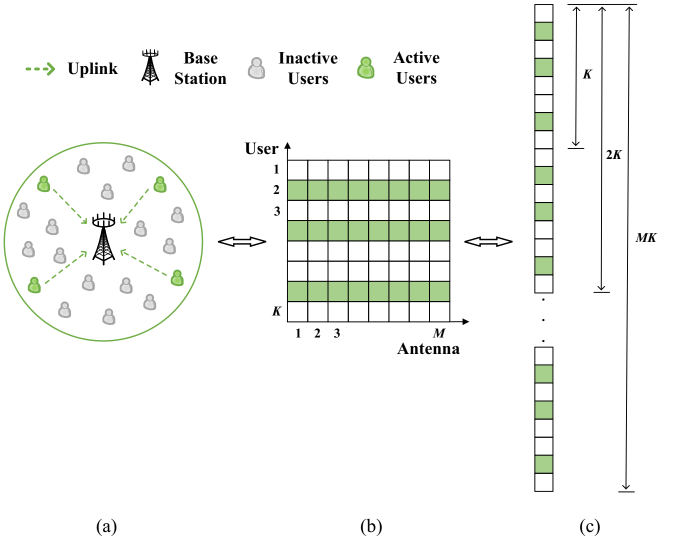

Fig. 1. (a) A typical uplink GF-NOMA system with multi-antenna reception; (b) spatial structure of user sparsity level; (c) block-wise SMV model.

transmit data to the BS, corresponding to the sparse nature of the data frame structure in mMTC. Specifically, the active user k chooses a symbol x_k from the finite alphabet $\aleph$ for data transmission, and conversely, if the k-th user is inactive, the transmitted symbol is equivalent to $x_k = 0$. Thus, the augmented alphabet set for all users can be represented by $\aleph_0 = \{\aleph\,, 0\}$. Due to the low power consumption of massive IoT applications, the binary phase shift keying (BPSK) modulation is adopted for symbol transmission, i.e., $\aleph_0 = \{-1, 0, 1\}$. Furthermore, each user k spreads its transmitted symbol x_k over a non-orthogonal sequence $\boldsymbol{\delta}_k = (\delta_{1k}, \delta_{2k}, \cdots, \delta_{Nk})^T$ of length N. Without loss of generality, we assume that the channel estimation and frame synchronization are completed before [5], [8]. Thus, the observed signal at the m-th antenna can be written as

$$\boldsymbol{r}^m = \sum_{k=1}^{K}(\boldsymbol{\varphi}_k^m \odot \boldsymbol{\delta}_k)x_k + \boldsymbol{n}^m = \boldsymbol{A}^m\boldsymbol{x} + \boldsymbol{n}^m, \tag{1}$$

where $\boldsymbol{\varphi}_k^m = (\varphi_{1k}^m, \varphi_{2k}^m, \cdots, \varphi_{Nk}^m)^T$ represents the channel fading coefficients between user k and the m-th antenna, $\boldsymbol{n}^m$ is the complex Gaussian noise with zero mean and $\sigma^2\boldsymbol{I}$ covariance, $\boldsymbol{A}^m = (\boldsymbol{\varphi}_1^m \odot \boldsymbol{\delta}_1, \boldsymbol{\varphi}_2^m \odot \boldsymbol{\delta}_2, \cdots, \boldsymbol{\varphi}_K^m \odot \boldsymbol{\delta}_K)$ is the equivalent measurement matrix, $\boldsymbol{x} = (x_1, x_2, \cdots, x_K)^T$ is the transmitted symbol for all users, and $\odot$ is the element-wise product.

Although the channel gains from each user to the BS are distinct, the sparse patterns of user activity are identical over multiple receiving antennas, as shown in Fig. 1(b). Based on this structured sparsity properties, (1) is extended into the block sparse single measurement vector (SMV) model for the MIMO case,

$$\mathbf{r} = \mathbf{A}\mathbf{s} + \mathbf{n}, \tag{2}$$

where $\mathbf{r} = \mathrm{vec}(\boldsymbol{r}^1, \boldsymbol{r}^2, \cdots, \boldsymbol{r}^M)$ is the linear received signal, $\mathbf{A} = \mathrm{diag}(\boldsymbol{A}^1, \boldsymbol{A}^2, \cdots, \boldsymbol{A}^M)$ is the block sensing matrix, $\mathbf{s} = \mathrm{vec}(\boldsymbol{x}, \boldsymbol{x}, \cdots, \boldsymbol{x})$ is the block sparse signal, $\mathbf{n} = \mathrm{vec}(\boldsymbol{n}^1, \boldsymbol{n}^2, \cdots, \boldsymbol{n}^M)$ is the linear Gaussian noise, and $\mathrm{vec}(\cdot)$ is the column-wise vectorization.

The first priority of this paper is joint user activity and data detection, i.e., recovering the transmitted symbol $\boldsymbol{x}$ from the block sparse vector $\mathbf{s}$ using the block CS theory. $\mathbf{s}$ is composed of M sparse vectors in which the non-zero elements have the same sparse location, as depicted in Fig. 1(c). However, this block sparse structure enables further improvement of MUD accuracy.

III. SE-VBI for MUD

SE-SBL is a robust sparse reconstruction algorithm for grant-free MIMO-NOMA systems that exploits both the spatial correlation and the sparsity of user activity. However, the SE-SBL algorithm ignores the finite alphabet constraint, i.e., all symbols transmitted by users are discrete random variables from the augmented alphabet $\aleph_0$. To this end, we introduce a Gaussian mixture model (GMM) [10] to capture the discreteness and sparsity of the transmitted signal, and use the variational Bayesian inference (VBI) method [11] to approximate the posterior probability distributions of the hidden variables. The proposed MUD algorithm, named spatially enhanced VBI (SE-VBI), is also available for MIMO-enabled GF-NOMA system by exploiting the diversity combining technique.

SE-VBI adopts a three-layer hierarchical prior model with sparsity promotion and finite alphabet constraint. Specifically, in the first layer, a multivariate Gaussian mixing prior is imposed on $\mathbf{s}$, i.e.,

$$\begin{aligned} p(\mathbf{s}\,;\boldsymbol{\alpha},\boldsymbol{z}) = &\prod_{m=1}^{M}\prod_{k=1}^{K}\mathcal{CN}(s_{mk}\mid -1\,,\alpha_{mk,1}^{-1})^{z_{mk,1}} \\ &\times \mathcal{CN}(s_{mk}\mid 0\,,\alpha_{mk,2}^{-1})^{z_{mk,2}}\mathcal{CN}(s_{mk}\mid 1\,,\alpha_{mk,3}^{-1})^{z_{mk,3}}, \end{aligned} \tag{3}$$

where s_{mk} is the mk-th element of the sparse vector $\mathbf{s}$, and $\{-1, 0, 1\}$, $\{\alpha_{mk,1}^{-1}, \alpha_{mk,2}^{-1}, \alpha_{mk,3}^{-1}\}$, and $\{z_{mk,1}, z_{mk,2}, z_{mk,3}\}$ are the mean, variance, and indicator factor for all Gaussian probability density components, respectively. Here, $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3\}$ with $\boldsymbol{\alpha}_l = (\alpha_{11,l}, \cdots, \alpha_{mk,l}, \cdots, \alpha_{MK,l})^T$, $\boldsymbol{z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3\}$ with $\boldsymbol{z}_l = (z_{11,l}, \cdots, z_{mk,l}, \cdots, z_{MK,l})^T$, and $l = 1, 2, 3$. Furthermore, we impose some restrictions on the indicator hyperparameters, i.e.,

$$z_{mk,1} + z_{mk,2} + z_{mk,3} = 1, \tag{4}$$

and

$$\{z_{mk,1}, z_{mk,2}, z_{mk,3}\} = \begin{cases} \{1, 0, 0\}, & s_{mk} \to -1, \\ \{0, 1, 0\}, & s_{mk} \to 0, \\ \{0, 0, 1\}, & s_{mk} \to 1. \end{cases} \tag{5}$$

The second layer assigns a multivariate Gamma prior to the hyperparameters $\boldsymbol{\alpha}$,

$$\begin{aligned} p(\boldsymbol{\alpha}) = &\prod_{m=1}^{M}\prod_{k=1}^{K}\mathrm{Gamma}(\alpha_{mk,1}\mid a\,, b) \\ &\times \mathrm{Gamma}(\alpha_{mk,2}\mid a\,, b)\mathrm{Gamma}(\alpha_{mk,3}\mid a\,, b), \end{aligned} \tag{6}$$

where a and b are informative parameters used to promote the sparsity of the transmitted signal $\mathbf{s}$. Meanwhile, in the second layer, the hyperparameters $\boldsymbol{z}$ are formulated as a multivariate polynomial distribution as follows

$$p(\boldsymbol{z}\,;\boldsymbol{\rho}) = \prod_{m=1}^{M}\prod_{k=1}^{K} \rho_{mk,1}^{z_{mk,1}}, \rho_{mk,2}^{z_{mk,2}}, \rho_{mk,3}^{z_{mk,3}}, \tag{7}$$

where $\{\rho_{mk,1}, \rho_{mk,2}, \rho_{mk,3}\}$ are the percentage coefficients of the GMM components, $\boldsymbol{\rho} = \{\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \boldsymbol{\rho}_3\}$ with $\boldsymbol{\rho}_l = (\rho_{11,l}, \cdots, \rho_{mk,l}, \cdots, \rho_{MK,l})^T$, and $l = 1, 2, 3$.

The third level specifies a multivariate Dirichlet prior to the hyperparameters $\boldsymbol{\rho}$,

$$p(\boldsymbol{\rho}) = \prod_{m=1}^{M}\prod_{k=1}^{K} \rho_{mk,1}^{\theta_1 - 1}\rho_{mk,2}^{\theta_2 - 1}\rho_{mk,3}^{\theta_3 - 1}, \tag{8}$$

where $\{\theta_1, \theta_2, \theta_3\}$ are constant parameters and the percentage hyperparameters satisfy

$$\rho_{mk,1} + \rho_{mk,2} + \rho_{mk,3} = 1,\ \rho_{mk,l} \in (0, 1). \tag{9}$$

For simplicity, $\gamma = 1/\sigma^2$ is defined as the noise precision and the noise vector is formulated as a multivariate Gaussian distribution, i.e., $p(\mathbf{n}) = \mathcal{CN}(\mathbf{n} \mid 0\,, \gamma^{-1}\boldsymbol{I})$. Furthermore, γ is interpreted as a hyperprior $p(\gamma) = \text{Gamma}(\gamma \mid c\,, d)$, where c and d are non-informative parameters. Thus, the likelihood distribution of the observed signal $\mathbf{r}$ is given by

$$p(\mathbf{r} \mid \mathbf{s}\,;\gamma) = \mathcal{CN}(\mathbf{r} \mid \mathbf{A}\mathbf{s}\,, \gamma^{-1}\boldsymbol{I}). \tag{10}$$

In the following, we focus on the estimation problem of the hidden variables $\boldsymbol{\Theta} = \{\mathbf{s}, \boldsymbol{\alpha}, \mathbf{z}, \boldsymbol{\rho}, \gamma\}$. Based on the three-layer hierarchical prior model, the complete likelihood distribution can be expressed as

$$p(\mathbf{r}\,;\boldsymbol{\Theta}) = p(\mathbf{r} \mid \mathbf{s}\,;\gamma)p(\mathbf{s}\,;\boldsymbol{\alpha}, \boldsymbol{z})p(\boldsymbol{\alpha})p(\boldsymbol{z}\,;\boldsymbol{\rho})p(\boldsymbol{\rho})p(\gamma). \tag{11}$$

The solution of $\boldsymbol{\Theta}$ can be obtained by calculating the maximum a posteriori (MAP) estimation of $p(\boldsymbol{\Theta} \mid \mathbf{r})$, which depends on the marginalized integration $\int p(\mathbf{r}\,;\boldsymbol{\Theta})d\boldsymbol{\Theta}$. Thus, the direct calculation of MAP is extremely complicated and it is also difficult to optimize it. For this purpose, we employ the VBI method to approximate the true posterior distribution instead of computing the MAP explicitly. The joint posterior probability can be approximated and factorized as

$$\begin{aligned} p(\boldsymbol{\Theta} \mid \mathbf{r}) &\approx q(\boldsymbol{\Theta}) \\ &= q(\mathbf{s})q(\boldsymbol{\alpha})q(\boldsymbol{z})q(\boldsymbol{\rho})q(\gamma), \end{aligned} \tag{12}$$

where $q(\boldsymbol{\alpha}) = q(\boldsymbol{\alpha}_1)q(\boldsymbol{\alpha}_2)q(\boldsymbol{\alpha}_3)$, $q(\boldsymbol{z}) = q(\boldsymbol{z}_1)q(\boldsymbol{z}_2)q(\boldsymbol{z}_3)$, and $q(\boldsymbol{\rho}) = q(\boldsymbol{\rho}_1)q(\boldsymbol{\rho}_2)q(\boldsymbol{\rho}_3)$.

According to [11], the optimal solution of (12) can be given by the following equation

$$\begin{aligned} \ln q(\Theta_i) =& \langle \ln p(\mathbf{r}\,;\boldsymbol{\Theta})\rangle_{\prod_{j\neq i} q(\Theta_j)} + \text{const}, \\ & i, j = 1, 2, 3, 4, 5, \end{aligned} \tag{13}$$

where Θ_i denotes the i-th hidden variable in $\boldsymbol{\Theta}$ and $\langle p(\mathbf{r}\,;\boldsymbol{\Theta})\rangle_{\prod_{j\neq i} q(\Theta_j)}$ represents the expectation for all variables except Θ_i. More precisely, the elaborated update steps are provided as follows.

1) Update of $q(\mathbf{s})$:

$$\begin{aligned} &\ln q(\mathbf{s}) \\ &= \langle \ln p(\mathbf{r}\,;\boldsymbol{\Theta})\rangle_{q(\boldsymbol{\alpha})q(\boldsymbol{z})q(\boldsymbol{\rho})q(\gamma)} + \text{const} \\ &= -\langle\gamma\rangle\|\mathbf{r} - \mathbf{A}\mathbf{s}\|_2^2 - \sum_{m=1}^{M}\sum_{k=1}^{K}[\langle z_{mk,1}\rangle\langle\alpha_{mk,1}\rangle \\ &\quad \times (s_{mk} + 1)^2 + \langle z_{mk,2}\rangle\langle\alpha_{mk,2}\rangle s_{mk}^2 \\ &\quad + \langle z_{mk,3}\rangle\langle\alpha_{mk,3}\rangle(s_{mk} - 1)^2] + \text{const} \\ &= -\mathbf{s}^H[\langle\gamma\rangle\mathbf{A}^H\mathbf{A} + \boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_2 + \boldsymbol{\Gamma}_3]\mathbf{s} \\ &\quad + 2(\langle\gamma\rangle\mathbf{r}^H\mathbf{A} - \boldsymbol{\Lambda}_1 + \boldsymbol{\Lambda}_3)\mathbf{s} + \text{const}, \end{aligned} \tag{14}$$

where

$$\begin{aligned} \boldsymbol{\Gamma}_1 &= \text{diag}(\langle z_{11,1}\rangle\langle\alpha_{11,1}\rangle, \cdots, \langle z_{MK,1}\rangle\langle\alpha_{MK,1}\rangle), \\ \boldsymbol{\Gamma}_2 &= \text{diag}(\langle z_{11,2}\rangle\langle\alpha_{11,2}\rangle, \cdots, \langle z_{MK,2}\rangle\langle\alpha_{MK,2}\rangle), \\ \boldsymbol{\Gamma}_3 &= \text{diag}(\langle z_{11,3}\rangle\langle\alpha_{11,3}\rangle, \cdots, \langle z_{MK,3}\rangle\langle\alpha_{MK,3}\rangle), \end{aligned} \tag{15}$$

and

$$\begin{aligned} \boldsymbol{\Lambda}_1 &= (\langle z_{11,1}\rangle\langle\alpha_{11,1}\rangle, \cdots, \langle z_{MK,1}\rangle\langle\alpha_{MK,1}\rangle), \\ \boldsymbol{\Lambda}_3 &= (\langle z_{11,3}\rangle\langle\alpha_{11,3}\rangle, \cdots, \langle z_{MK,3}\rangle\langle\alpha_{MK,3}\rangle). \end{aligned} \tag{16}$$

It is evident from (14) that $q(\mathbf{s})$ satisfies a Gaussian distribution, i.e.,

$$q(\mathbf{s}) = \mathcal{CN}(\mathbf{s} \mid \boldsymbol{\mu}\,, \boldsymbol{\Sigma}), \tag{17}$$

where the posterior mean and covariance can be formulated as

$$\mathbf{s}^{\text{VBI}} = \boldsymbol{\mu} = \boldsymbol{\Sigma}(\langle\gamma\rangle\mathbf{r}^H\mathbf{A} - \boldsymbol{\Lambda}_1 + \boldsymbol{\Lambda}_3)^H, \tag{18}$$

and

$$\boldsymbol{\Sigma} = (\langle\gamma\rangle\mathbf{A}^H\mathbf{A} + \boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_2 + \boldsymbol{\Gamma}_3)^{-1}. \tag{19}$$

The mean is treated as the minimum mean square error (MMSE) estimation of the block sparse vector. Then, we can obtain M observed samples of the transmitted signal $\boldsymbol{x}$ by the inverse vectorization,

$$\boldsymbol{S} = (\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_M) = \text{vec}^{-1}(\mathbf{s}^{\text{VBI}}, M). \tag{20}$$

Given the space diversity of multi-antenna reception and the block sparsity of user activity, the multiple observed signals are combined as follows

$$\hat{\boldsymbol{x}} = \sum_{m=1}^{M} \eta_m \boldsymbol{S}_m, \tag{21}$$

where η_m is the diversity gain of the m-th antenna. For simplicity, the equal gain combining (EGC) technique [12] is adopted for multiple observed samples, i.e., $\eta = 1/M$. Furthermore, we extend the combined signal into a block pattern to improve the posterior mean of $\mathbf{s}$,

$$\mathbf{s}^{\text{SE}-\text{VBI}} = \hat{\boldsymbol{\mu}} = \text{vec}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{x}}, \cdots, \hat{\boldsymbol{x}}). \tag{22}$$

Next, the remaining hidden variables are updated using the improved posterior distribution $q(\mathbf{s}) = \mathcal{CN}(\mathbf{s} \mid \hat{\boldsymbol{\mu}}\,, \boldsymbol{\Sigma})$.

2) Update of $q(\boldsymbol{\alpha})$:

$$\begin{aligned}
&\ln q(\boldsymbol{\alpha})\\
&= \langle \ln p(\mathbf{r}\,;\boldsymbol{\Theta})\rangle_{q(\mathbf{s})q(\boldsymbol{z})q(\boldsymbol{\rho})q(\gamma)} + \text{const}\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K}[\langle z_{mk,1}\rangle \ln \alpha_{mk,1} - \langle (s_{mk}+1)^2\rangle \langle z_{mk,1}\rangle \alpha_{mk,1}\\
&\quad + \langle z_{mk,2}\rangle \ln \alpha_{mk,2} - \langle s_{mk}^2\rangle \langle z_{mk,2}\rangle \alpha_{mk,2}\\
&\quad + \langle z_{mk,3}\rangle \ln \alpha_{mk,3} - \langle (s_{mk}-1)^2\rangle \langle z_{mk,3}\rangle \alpha_{mk,3}]\\
&\quad + \sum_{m=1}^{M}\sum_{k=1}^{K}[(a-1)\ln \alpha_{mk,1} - b\alpha_{mk,1} + (a-1)\ln \alpha_{mk,2}\\
&\quad - b\alpha_{mk,2} + (a-1)\ln \alpha_{mk,3} - b\alpha_{mk,3}] + \text{const}.
\end{aligned} \tag{23}$$

Since $\ln q(\boldsymbol{\alpha}) = \sum_{l=1}^{3} \ln q(\boldsymbol{\alpha}_l)$, we have

$$\begin{aligned}
\ln q(\boldsymbol{\alpha}_1) = &\sum_{m=1}^{M}\sum_{k=1}^{K}[\langle z_{mk,1}\rangle \ln \alpha_{mk,1} - \langle (s_{mk}+1)^2\rangle \langle z_{mk,1}\rangle\\
&\times \alpha_{mk,1} + (a-1)\ln \alpha_{mk,1} - b\alpha_{mk,1}] + \text{const}.
\end{aligned} \tag{24}$$

Each term in (24) can be written as

$$\begin{aligned}
\ln q(\alpha_{mk,1}) =&(a + \langle z_{mk,1}\rangle - 1)\ln \alpha_{mk,1}\\
&- (b + \langle (s_{mk}+1)^2\rangle \langle z_{mk,1}\rangle)\alpha_{mk,1} + \text{const}.
\end{aligned} \tag{25}$$

Hence, $q(\alpha_{mk,1})$ obeys a parameterized Gamma distribution, i.e.,

$$q(\alpha_{mk,1}) = \text{Gamma}(\alpha_{mk,1} \mid \hat{a}_{mk,1}\,, \hat{b}_{mk,1}), \tag{26}$$

where the parameters $\hat{a}_{mk,1}$ and $\hat{b}_{mk,1}$ are as follows

$$\begin{aligned}
\hat{a}_{mk,1} &= a + \langle z_{mk,1}\rangle,\\
\hat{b}_{mk,1} &= b + \langle (s_{mk}+1)^2\rangle \langle z_{mk,1}\rangle.
\end{aligned} \tag{27}$$

Analogously, $q(\alpha_{mk,2})$ and $q(\alpha_{mk,3})$ can be generalized to the Gamma distribution,

$$q(\alpha_{mk,2}) = \text{Gamma}(\alpha_{mk,2} \mid \hat{a}_{mk,2}\,, \hat{b}_{mk,2}), \tag{28}$$

where the parameters $\hat{a}_{mk,2}$ and $\hat{b}_{mk,2}$ satisfy

$$\begin{aligned}
\hat{a}_{mk,2} &= a + \langle z_{mk,2}\rangle,\\
\hat{b}_{mk,2} &= b + \langle s_{mk}^2\rangle \langle z_{mk,2}\rangle,
\end{aligned} \tag{29}$$

and

$$q(\alpha_{mk,3}) = \text{Gamma}(\alpha_{mk,3} \mid \hat{a}_{mk,3}\,, \hat{b}_{mk,3}), \tag{30}$$

where the parameters $\hat{a}_{mk,3}$ and $\hat{b}_{mk,3}$ satisfy

$$\begin{aligned}
\hat{a}_{mk,3} &= a + \langle z_{mk,3}\rangle,\\
\hat{b}_{mk,3} &= b + \langle (s_{mk}-1)^2\rangle \langle z_{mk,3}\rangle.
\end{aligned} \tag{31}$$

The expectation of $\alpha_{mk,l}$ and $\ln \alpha_{mk,l}$ can be expressed as

$$\langle \alpha_{mk,l}\rangle = \frac{\hat{a}_{mk,l}}{\hat{b}_{mk,l}},\ l = 1, 2, 3, \tag{32}$$

and

$$\langle \ln \alpha_{mk,l}\rangle = \Psi(\hat{a}_{mk,l}) - \ln \hat{b}_{mk,l},\ l = 1, 2, 3, \tag{33}$$

where $\Psi(\cdot)$ denotes the digamma function.

3) Update of $q(\boldsymbol{z})$:

$$\begin{aligned}
&\ln q(\boldsymbol{z})\\
&= \langle \ln p(\mathbf{r}\,;\boldsymbol{\Theta})\rangle_{q(\mathbf{s})q(\boldsymbol{\alpha})q(\boldsymbol{\rho})q(\gamma)} + \text{const}\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K}[\langle \ln \alpha_{mk,1}\rangle z_{mk,1} - \langle (s_{mk}+1)^2\rangle \langle \alpha_{mk,1}\rangle z_{mk,1}\\
&\quad + \langle \ln \alpha_{mk,2}\rangle z_{mk,2} - \langle s_{mk}^2\rangle \langle \alpha_{mk,2}\rangle z_{mk,2}\\
&\quad + \langle \ln \alpha_{mk,3}\rangle z_{mk,3} - \langle (s_{mk}-1)^2\rangle \langle \alpha_{mk,3}\rangle z_{mk,3}]\\
&\quad + \sum_{m=1}^{M}\sum_{k=1}^{K}[\langle \ln \rho_{mk,1}\rangle z_{mk,1} + \langle \ln \rho_{mk,2}\rangle z_{mk,2}\\
&\quad + \langle \ln \rho_{mk,3}\rangle z_{mk,3}] + \text{const}.
\end{aligned} \tag{34}$$

Since $\ln q(\boldsymbol{z}) = \sum_{l=1}^{3} \ln q(\boldsymbol{z}_l)$, we have

$$\begin{aligned}
\ln q(\boldsymbol{z}_1) = &\sum_{m=1}^{M}\sum_{k=1}^{K}[\langle \ln \alpha_{mk,1}\rangle z_{mk,1} - \langle (s_{mk}+1)^2\rangle \langle \alpha_{mk,1}\rangle\\
&\times z_{mk,1} + \langle \ln \rho_{mk,1}\rangle z_{mk,1}] + \text{const}.
\end{aligned} \tag{35}$$

Each term in (35) can be written as

$$\begin{aligned}
\ln q(z_{mk,1}) =&(\langle \ln \alpha_{mk,1}\rangle - \langle (s_{mk}+1)^2\rangle \langle \alpha_{mk,1}\rangle\\
&+ \langle \ln \rho_{mk,1}\rangle) z_{mk,1} + \text{const}.
\end{aligned} \tag{36}$$

Similarly, we can obtain the logarithmic posterior probabilities of $z_{mk,2}$ and $z_{mk,3}$,

$$\begin{aligned}
\ln q(z_{mk,2}) =&(\langle \ln \alpha_{mk,2}\rangle - \langle s_{mk}^2\rangle \langle \alpha_{mk,2}\rangle\\
&+ \langle \ln \rho_{mk,2}\rangle) z_{mk,2} + \text{const},
\end{aligned} \tag{37}$$

and

$$\begin{aligned}
\ln q(z_{mk,3}) =&(\langle \ln \alpha_{mk,3}\rangle - \langle (s_{mk}-1)^2\rangle \langle \alpha_{mk,3}\rangle\\
&+ \langle \ln \rho_{mk,3}\rangle) z_{mk,3} + \text{const}.
\end{aligned} \tag{38}$$

Under the setting $\sum_{l=1}^{3} z_{mk,l} = 1$, we define the expectation of $z_{mk,l}$ as

$$\begin{aligned}
\langle z_{mk,1}\rangle &= \frac{P_{mk,1}}{P_{mk,1} + P_{mk,2} + P_{mk,3}},\\
\langle z_{mk,2}\rangle &= \frac{P_{mk,2}}{P_{mk,1} + P_{mk,2} + P_{mk,3}},\\
\langle z_{mk,3}\rangle &= \frac{P_{mk,3}}{P_{mk,1} + P_{mk,2} + P_{mk,3}},
\end{aligned} \tag{39}$$

where

$$\begin{aligned}
P_{mk,1} &= \exp(\langle \ln \alpha_{mk,1}\rangle - \langle (s_{mk}+1)^2\rangle \langle \alpha_{mk,1}\rangle + \langle \ln \rho_{mk,1}\rangle),\\
P_{mk,2} &= \exp(\langle \ln \alpha_{mk,2}\rangle - \langle s_{mk}^2\rangle \langle \alpha_{mk,2}\rangle + \langle \ln \rho_{mk,2}\rangle),\\
P_{mk,3} &= \exp(\langle \ln \alpha_{mk,3}\rangle - \langle (s_{mk}-1)^2\rangle \langle \alpha_{mk,3}\rangle + \langle \ln \rho_{mk,3}\rangle).
\end{aligned} \tag{40}$$

4) Update of $q(\boldsymbol{\rho})$:

$$\begin{aligned}
&\ln q(\boldsymbol{\rho})\\
&= \langle \ln p(\mathbf{r}\,;\boldsymbol{\Theta})\rangle_{q(\mathbf{s})q(\boldsymbol{\alpha})q(\boldsymbol{z})q(\gamma)} + \text{const}\\
&= \sum_{m=1}^{M}\sum_{k=1}^{K}[\langle z_{mk,1}\rangle \ln \rho_{mk,1} + \langle z_{mk,2}\rangle \ln \rho_{mk,2}\\
&\quad + \langle z_{mk,3}\rangle \ln \rho_{mk,3}] + \sum_{m=1}^{M}\sum_{k=1}^{K}[(\theta_1 - 1)\ln \rho_{mk,1}\\
&\quad + (\theta_2 - 1)\ln \rho_{mk,2} + (\theta_3 - 1)\ln \rho_{mk,3}] + \text{const}.
\end{aligned} \tag{41}$$

Since $\ln q(\boldsymbol{\rho}) = \sum_{l=1}^{3} \ln q(\boldsymbol{\rho}_l)$, we have

$$\begin{aligned}
\ln q(\boldsymbol{\rho}_1) = \sum_{m=1}^{M}\sum_{k=1}^{K}&[\langle z_{mk,1}\rangle \ln \rho_{mk,1}\\
&+ (\theta_1 - 1)\ln \rho_{mk,1}] + \text{const}.
\end{aligned} \tag{42}$$

Each term in (42) can be written as

$$\ln q(\rho_{mk,1}) = (\theta_1 + \langle z_{mk,1}\rangle - 1)\ln \rho_{mk,1} + \text{const}. \tag{43}$$

Moreover, the log-posterior probabilities of $q(\rho_{mk,2})$ and $q(\rho_{mk,3})$ are given by

$$\ln q(\rho_{mk,2}) = (\theta_2 + \langle z_{mk,2}\rangle - 1)\ln \rho_{mk,2} + \text{const}, \tag{44}$$

and

$$\ln q(\rho_{mk,3}) = (\theta_3 + \langle z_{mk,3}\rangle - 1)\ln \rho_{mk,3} + \text{const}. \tag{45}$$

Given that $\sum_{l=1}^{3} \rho_{mk,l} = 1$, we define the expectation of $\ln \rho_{mk,l}$ as

$$\langle \ln \rho_{mk,l}\rangle = \Psi(Q_{mk,l}) - \Psi(\sum_{l=1}^{3} Q_{mk,l}), \tag{46}$$

where

$$Q_{mk,l} = \theta_l + \langle z_{mk,l}\rangle. \tag{47}$$

5) Update of $q(\gamma)$:

$$\begin{aligned}
&\ln q(\gamma)\\
&= \langle \ln p(\mathbf{r}\,;\boldsymbol{\Theta})\rangle_{q(\mathbf{s})q(\boldsymbol{\alpha})q(\boldsymbol{z})q(\boldsymbol{\rho})} + \text{const}\\
&= NM \ln \gamma - \langle \|\mathbf{r} - \mathbf{A}\mathbf{s}\|_2^2\rangle \gamma + (c-1)\ln \gamma - d\gamma + \text{const}\\
&= (c + NM - 1)\ln \gamma - (d + \langle \|\mathbf{r} - \mathbf{A}\mathbf{s}\|_2^2\rangle)\gamma + \text{const}.
\end{aligned} \tag{48}$$

It is observed that $q(\gamma)$ obeys a Gamma distribution, i.e.,

$$q(\gamma) = \text{Gamma}(\gamma \mid \hat{c}\,, \hat{d}), \tag{49}$$

where the parameters $\hat{c}$ and $\hat{d}$ are denoted as

$$\begin{aligned}
\hat{c} &= c + NM,\\
\hat{d} &= d + \langle \|\mathbf{r} - \mathbf{A}\mathbf{s}\|_2^2\rangle.
\end{aligned} \tag{50}$$

Based on this, the expectation of γ can be calculated as

$$\langle \gamma \rangle = \frac{\hat{c}}{\hat{d}} = \frac{c + NM}{d + \|\mathbf{r} - \mathbf{A}\hat{\boldsymbol{\mu}}\|_2^2 + \text{tr}[\mathbf{A}^H\mathbf{A}\boldsymbol{\Sigma}]}. \tag{51}$$

IV. SE-CoFe-VBI for MUD

SE-VBI is an improved sparse detection method for multi-antenna reception. Unfortunately, inferring the posterior covariance of $\mathbf{s}$ requires a high-dimensional matrix inversion, as mentioned in (19). The computational cost of SE-VBI is $\mathcal{O}(K^3M^3)$ per iteration, which is burdensome and prohibitive for massive connectivity. To this end, we introduce the CoFe strategy [9] into the SE-VBI framework to produce a cost-effective MUD method called SE-CoFe-VBI. Recalling the steps of SE-VBI, $\boldsymbol{\Sigma}$ is needed to update $\boldsymbol{\mu}$ in (18), $\langle \alpha_{mk,l}\rangle$ in (32), $\langle \ln \alpha_{mk,l}\rangle$ in (33), $\langle z_{mk,l}\rangle$ in (39), and $\langle \gamma \rangle$ in (51). CoFe provides a streamlined update of the above expectations without explicitly calculating $\boldsymbol{\Sigma}$, using the diagonal estimation principle and fundamental matrix theory.

1) Update of $\boldsymbol{\mu}$: After a basic matrix transformation, (18) can be re-written as

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = (\langle \gamma \rangle \mathbf{r}^H\mathbf{A} - \boldsymbol{\Lambda}_1 + \boldsymbol{\Lambda}_3)^H. \tag{52}$$

As such, the posterior mean can be estimated by solving the linear equation $\boldsymbol{\Phi}\mathbf{v}_\mu = \mathbf{u}_\mu$ for $\boldsymbol{\Phi} = (\langle \gamma \rangle \mathbf{A}^H\mathbf{A} + \boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_2 + \boldsymbol{\Gamma}_3)$, $\mathbf{v}_\mu = \boldsymbol{\mu}$, and $\mathbf{u}_\mu = (\langle \gamma \rangle \mathbf{r}^H\mathbf{A} - \boldsymbol{\Lambda}_1 + \boldsymbol{\Lambda}_3)^H$.

2) Update of $\boldsymbol{\Sigma}$: As seen in (32), (33), and (39), only the diagonal elements of $\boldsymbol{\Sigma}$ are required to update these hyperparameters without computing the entire matrix. Therefore, we aim to estimate the diagonal entries of $\boldsymbol{\Sigma}$ by the following diagonal estimation rule [9].

Lemma 1: We define $\mathbf{H} \in \mathbb{C}^{G\times G}$ and $\mathbf{h} \in \mathbb{C}^{G\times 1}$ as arbitrary square matrix and its diagonal vector. The unbiased estimation of $\mathbf{h}$ can be expressed as

$$\mathbf{h} = (\sum_{w=1}^{W} \mathbf{u}_w \odot \mathbf{H}\mathbf{u}_w) \oslash (\sum_{w=1}^{W} \mathbf{u}_w \odot \mathbf{u}_w), \tag{53}$$

where $\mathbf{u}_w \in \mathbb{C}^{G\times 1}$, $\forall w \in \{1, 2, \cdots, W\}$ are random probe vectors satisfying $\langle \mathbf{u}_w \rangle = 0$ and $\oslash$ is the element-wise division. For simplicity, we assume that the stochastic vectors obey the Randemacher distribution, i.e., $\mathbf{u}_w$ draws values from $\{-1, 1\}$ with equal probability. Thus, the diagonal estimator can be reduced to

$$\mathbf{h} = \frac{1}{W}\sum_{w=1}^{W} \mathbf{u}_w \odot \mathbf{H}\mathbf{u}_w. \tag{54}$$

Thanks to the above lemma, the diagonal vector estimation problem for $\boldsymbol{\Sigma}$ is transformed into solving a system of linear equations $\boldsymbol{\Phi}[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_W] = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_W]$. The solution $\mathbf{v}_w$ is the estimated $\boldsymbol{\Sigma}\mathbf{u}_w$ used in (54) to compute the diagonal elements.

3) Update of $\langle \gamma \rangle$: It is obvious from (51) that the diagonal elements of $\mathbf{A}^H\mathbf{A}\boldsymbol{\Sigma}$ are required for updating the posterior distribution of γ. This motivates use of the diagonal estimation rule to calculate $\text{diag}(\mathbf{A}^H\mathbf{A}\boldsymbol{\Sigma})$. Based on the estimated $\boldsymbol{\Sigma}\mathbf{u}_w$, the diagonal vector can be expressed as

$$\text{diag}(\mathbf{A}^H\mathbf{A}\boldsymbol{\Sigma}) = \frac{1}{W}\sum_{w=1}^{W} \mathbf{u}_w \odot \mathbf{A}^H\mathbf{A}\boldsymbol{\Sigma}\mathbf{u}_w. \tag{55}$$

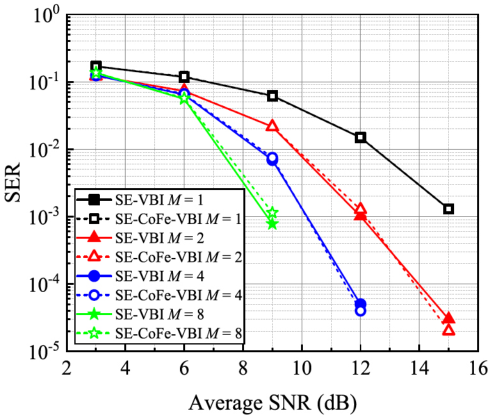

Fig. 2. SER performance of the proposed MUD algorithms for multi-antenna reception.

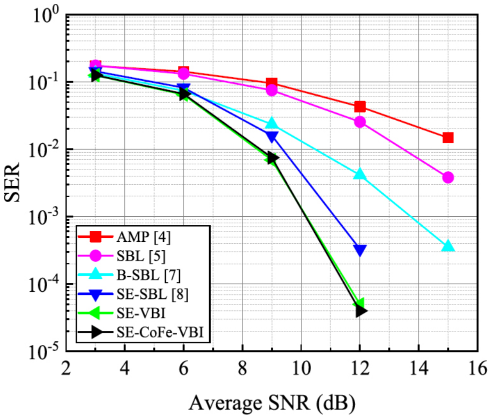

Fig. 3. SER comparison of several MUD algorithms for $M = 4$ antenna reception.

The parallel conjugate gradient (PCG) algorithm [9] is adopted to solve these $W + 1$ linear equations. Since the complexity of PCG depends on matrix-matrix multiplication, the computational cost of SE-CoFe-VBI is $\mathcal{O}(K^2M^2(W+1))$ for single iteration.

V. Simulation Results

In this section, experimental results are presented to validate the proposed MUD algorithms. Specifically, the proposed SE-VBI and SE-CoFe-VBI algorithms are evaluated and compared with the standard AMP [4], SBL [5], B-SBL [7], and SE-SBL [8] methods. We consider an overloaded mMTC scenario in which the total number of users is $K = 200$, the user sparsity level is $\epsilon = 0.2$, and the spreading sequence length is $N = 160$. In all simulations, BPSK modulation and flat Rayleigh fading channel model are adopted. The number of random probe vectors is $W = 20$, which is an empirical value according to [9]. Moreover, the static parameters are $a = 1$, $b = c = d = 10^{-8}$, and $\theta_1 = \theta_2 = \theta_3 = 1$.

Fig. 2 shows the symbol error rate (SER) performance of the SE-VBI and SE-CoFe-VBI algorithms for multi-antenna reception. The single-antenna case is further considered for comparison. It is noticed that the SER performance improves significantly with more receiving antennas. This suggests that the proposed algorithms exploit the spatial structure of user sparsity patterns. Furthermore, the performance of SE-VBI and SE-CoFe-VBI is virtually identical, because the estimated posterior distribution is roughly comparable to the results of variational inference.

Fig. 3 shows the SER comparison of several MUD algorithms. The performance of all algorithms improves as the signal-to-noise ratio (SNR) increases. We can observe that SE-VBI and SE-CoFe-VBI outperform the alternatives especially in the high SNR region. This is because the proposed algorithms exploit the multivariate nature of discreteness, sparsity, and spatial correlation of the transmitted signal. Moreover, the proposed algorithms still perform better than the standard SBL method even in the single-antenna case, as presented in Figs. 2 and 3.

VI. Conclusions

In this paper, we developed two VBI-based MUD algorithms to improve sparse detection and reduce computational complexity in an uplink MIMO-enabled GF-NOMA system. The proposed SE-VBI and SE-CoFe-VBI algorithms exploit the multivariate nature of BPSK modulation to be more suitable for energy-constrained mMTC scenarios. Simulation results showed that the proposed algorithms not only outperform SE-SBL for multi-antenna reception, but also surpass SBL for single-antenna reception.

References

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. A. Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615-636, Mar. 2021.

[2] M. Mohammadkarimi, O. A. Dobre, and M. Z. Win, "Massive uncoordinated multiple access for beyond 5G," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 2969-2986, May 2022

[3] Y. Mei et al., "Compressive sensing-based joint activity and data detection for grant-free massive IoT access," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1851-1869, Mar. 2022.

[4] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640-643, Mar. 2017.

[5] X. Zhang, P. Fan, J. Liu, and L. Hao, "Bayesian learning-based multiuser detection for grant-free NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6317-6328, Aug. 2022.

[6] L. Liu and W. Yu, "Massive connectivity with massive MIMO-Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933-2946, Jun. 2018.

[7] Y. Zhang, Q. Guo, Z. Wang, J. Xi, and N. Wu, "Block sparse Bayesian learning based joint user activity detection and channel estimation for grant-free NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9631-9640, Oct. 2018.

[8] B. Yang, X. Zhang, L. Hao, and G. K. Karagiannidis, "Improved Bayesian learning detectors for uplink grant-free MIMO-NOMA," *IEEE Wireless Commun. Lett.*, vol. 12, no. 12, pp. 2243-2247, Dec. 2023.

[9] A. Lin et al., "Covariance-free sparse Bayesian learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 3818-3831, 2022.

[10] X. Zhang, Y. C. Liang, and J. Fang, "Novel Bayesian inference algorithms for multiuser detection in M2M communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7833-7848, Sep. 2017.

[11] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131-146, Nov. 2008.

[12] F. Clazzer, C. Kissling, and M. Marchese, "Enhancing contention resolution ALOHA using combining techniques," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2576-2587, Jun. 2018.