Efficient Enhanced K-Means Clustering for Semi-Blind Channel Estimation of Cell-Free Massive MIMO

Xuefeng Huang[†], Xu Zhu[*], Yufei Jiang[†] and Yujie Liu[*]
† School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, China
* Department of Electrical Engineering and Electronics, The University of Liverpool, Liverpool, UK
Emails: 18s052129@stu.hit.edu.cn, xuzhu@liverpool.ac.uk, jiangyufei@hit.edu.cn, yujieliu@liverpool.ac.uk

***Abstract*— We propose an efficient enhanced K-means clustering (E-KMC) algorithm for semi-blind channel estimation of uplink cell-free massive multiple-input multiple-output (MIMO) systems in factory automation, an important application of the internet of things (IoT). The proposed E-KMC algorithm operates with significantly less clusters and complexity than the KMC algorithm while achieving enhanced bit error rate (BER) performance, as the latter converges extremely slowly even with just medium modulation order and a medium number of transmit antennas. A near-optimal short pilot is designed to assist clustering of the E-KMC based channel estimation scheme. The semi-blind receiver structure achieves a BER performance that is very close to the case with perfect channel state information (CSI), as well as a mean square error (MSE) of channel estimation that is very close to the theoretical lower bound derived in the paper. The proposed E-KMC based channel estimation scheme also significantly outperforms other types of semi-blind channel estimation approaches including second- and higher-order statistics based and machine learning based approaches, while at a much lower complexity. In addition, the E-KMC based channel estimation, is conducted at central processing unit (CPU) and avoids excessive fronthaul overhead due to exchange of the estimated CSI between access points (APs) and CPU.**

I. Introduction

Cell-free (CF) massive multiple-input multiple-output (MIMO) [1] has been introduced as a promising fifth generation (5G) and beyond 5G technique, due to its high energy efficiency, high quality of service and flexible and cost-effective infrastructure. CF massive MIMO is an eminently suitable solution for internet of things (IoT) applications such as factory automation [2], where closed-loop control sensors and acuators run periodic cycles while demanding ultra-reliable low-latency communications [3]. In such a system, it is important to acquire accurate channel state information (CSI) with scant fronthaul overhead and relatively short frame length.

Conventional channel estimation for CF massive MIMO systems is based on training symbols [4] [5]. Their accuracy is highly dependent on the number of training symbols and orthogonality between training symbols. When CSI is estimated at individual access points (APs) and then passed to central processing unit (CPU) for joint signal detection [4] [5], additional fronthaul overhead is required for exchange of the estimated CSI between APs and CPU. Independent component analysis (ICA), one of the blind source separation approaches, was employed in [6] and [7] for blind channel estimation and signal detection in massive MIMO systems, however, the ICA-based approaches [6] [7] suffer an error floor at high signal-to-noise ratio (SNR) due to inadequate ambiguity elimination.

Machine learning [8] has been applied to MIMO systems [9] [10] [11]. In [9], an expectation-maximization (EM) based semi-blind channel estimation approach was proposed, however, its complexity is relatively high, and its performance is limited by the initial channel estimation which requires a relatively long pilot with low spectral efficiency. K-means clustering (KMC) [12], which enables noise suppression, is one of the most popular clustering algorithms [13] due to its effectiveness and simplicity. KMC was employed in [10] for blind signal detection in space shift keying MIMO systems, where KMC was conducted multiple times to avoid error floor, requiring a high computational complexity. The work was extended in [11] to present a so called KMC-based blind signal detector for MIMO systems with spatial modulation. However, it was assumed in [11] that ambiguity was mitigated with ideal CSI obtained, which is not practical. Therefore, the approach in [11] is only pseudo blind. Also, KMC [10] [11] is not suitable for a CF massive MIMO system with relatively short frame length, high modulation order and large number of transmit antennas.

Motivated by the above open issues, we propose an efficient enhanced KMC (E-KMC) algorithm based semi-blind channel estimation scheme for a CF massive MIMO system in factory automation. The main contributions are summarized as follows.

- The proposed E-KMC algorithm achieves a significant complexity reduction over KMC [11], while outperforming KMC in terms of bit error rate (BER). This is because given the number of transmit antennas K and

This work was supported in part by the National Natural Science Foundation of China under Grant 61901138, in part by the Natural Science Foundation of Guangdong Province under Grants 2018A030313344 and 2018A030313298, and in part by the Guangdong Science and Technology Planning Project 2018B030322004.

modulation order M, E-KMC forms only K clusters of predefined size with all received signal vectors, and requires only one iteration to assign each cluster a group of received signal vectors nearest to the mean (centroid) of the cluster. While KMC [11] has M^K clusters to assign and has to conduct cluster assignment and means update iteratively. It does not converge even after 1000 iterations and presents a worse BER performance than E-KMC. Furthermore, clustering requires the number of observations (frame length in this use case) to be much larger than the total number of clusters. Hence, the proposed E-KMC algorithm is suitable for a wide range of number of transmit antennas, modulation order and frame length.

- A near-optimal short pilot is designed with low complexity, which helps to find the centroids of clusters and avoid ambiguity. The proposed E-KMC based channel estimation scheme enables a BER performance that is very close to the case with perfect CSI. It also outperforms other types of semi-blind channel estimation approaches [6] [9] [14] in terms of BER and mean square error (MSE) of channel estimation, and requires much lower complexity than the ICA [6] and EM [9] receivers with the same pilot length. A lower bound on the MSE of channel estimation by E-KMC is derived and shown to be very lose to simulation results.
- As the proposed E-KMC based channel estimation is conducted jointly at CPU, the fronthaul overhead is greatly reduced compared to [4] and [5], where channel estimation is conducted at individual APs and transmitted to the CPU.

II. System Model

In this paper, we consider a CF massive MIMO uplink system for factory automation with K single-antenna sensors and N distributed single-antenna APs, as illustrated in Fig. 1. The APs are connected to a CPU via fronthaul links. Joint signal detection is implemented at CPU, without the need of any information of channel.

Each transmit symbol is drawn from a constellation set $\mathbb{C} = \{c_1, c_2, \cdots, c_M\}$, where M is the modulation order. $\mathbf{X} = [\mathbf{x}[1], \cdots, \mathbf{x}[t], \cdots, \mathbf{x}[T]]$ of size $K \times T$ is the transmitted signal matrix, where $\mathbf{x}[t]$ of size $K \times 1$ as its t-th column is defined as the t-th symbol vector of K sensors within a frame. The first L symbols of $\mathbf{X}$ are exploited as pilots, the rest $(T-L)$ symbols are utilized for data transmission. Thus, $\mathbf{X}$ can be written as $\mathbf{X} = [\mathbf{X}_P \ \mathbf{X}_D]$, where $\mathbf{X}_P = [\mathbf{x}_p[1], \cdots, \mathbf{x}_p[l], \cdots, \mathbf{x}_p[L]]$ and $\mathbf{X}_D = [\mathbf{x}_d[L+1], \cdots, \mathbf{x}_d[u], \cdots, \mathbf{x}_d[T]]$ respectively denote the pilot matrix of size $K \times L$ and data matrix of size $K \times (T-L)$. Since closed-loop control sensors run periodic cycles in factory automation [3], several data vectors occur with a higher probability than others. For example, vector $\mathbf{f}$ takes

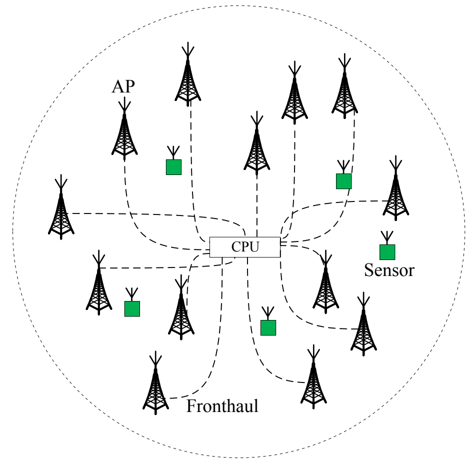

Fig. 1. Diagram of a CF massive MIMO system for factory automation

four possible values of $\mathbf{\Phi}_1$ to $\mathbf{\Phi}_4$ with unequal probabilities, where vectors $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ have a higher probability to occur than the $\mathbf{\Phi}_3$ and $\mathbf{\Phi}_4$.

The received signal at CPU with T samples is organized into an $N \times T$ matrix $\mathbf{Y}$, given by

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{W} \tag{1}$$

where $\mathbf{W}$ denotes the noise matrix of size $N \times T$, which has independent identically distributed entries following $\mathcal{CN}(0, N_0)$, which N_0 denoting the noise variance, and $\mathbf{H}$ is the Rayleigh block fading channel matrix. Let h_{nk} be the element of $\mathbf{H}$ in its n-th row and the k-th column, which represents the channel coefficient between the n-th AP and k-th sensor, which is modeled as $h_{nk} = g_{nk}\sqrt{\beta_{nk}}$ [1], where β_{nk} and $g_{nk} \in \mathcal{CN}(0,1)$ denote the large-scale fading and small-scale fading coefficients between the n-th AP and k-th sensor, respectively. $\mathbf{Y} = [\mathbf{y}[1], \cdots, \mathbf{y}[t], \cdots, \mathbf{y}[T]]$, where $\mathbf{y}[t]$ denotes the t-th column of $\mathbf{Y}$. The observed matrix at CPU is written as $\mathbf{Y} = [\mathbf{Y}_P \ \mathbf{Y}_D]$, where $\mathbf{Y}_P$ and $\mathbf{Y}_D$ correspond to the received pilot symbols and the received data symbols, respectively.

III. E-KMC Based Semi-Blind Channel Estimation

In this section, we present the proposed E-KMC based semi-blind channel estimator and the pilot design to enable high-accuracy and low-complexity estimation.

A. E-KMC Based Semi-blind Channel Estimation

To overcome the limitations of the pilot assisted approach and the conventional KMC method as stated above, we propose an E-KMC algorithm to estimate channel in a semi-blind manner with low complexity.

Let Ψ denote the set of all possible values of transmitted signal vectors with size Q. The pilot vector $\mathbf{x}_p[l]$ is drawn from $\Psi_P \in \Psi$ whose size is L, and the data vector $\mathbf{x}_d[u]$ is drawn from $\Psi_D = \Psi - \Psi_P$ with size $Q - L$. As stated in Section II, some data vectors are likely to appear with a higher probability in the constellation for factory automation. In E-KMC, the L data vectors appeared with highest probability which are determined by forecasting the distribution of all data vectors in advance, are respectively selected and encoded as the L pilot vectors, and the corresponding decoding process is performed after signal detection.

Let $\mathbf{x}_{cd}[u]$ denote the u-th coded data vector. After the coding process, the number of the coded data vectors $\mathbf{x}_{cd}[u], u = L+1, \cdots, T$ equal to each column vector of pilot matrix $\mathbf{x}_p[l]$, will be relatively large in factory automation and we denote it as T_c^l. For convenience, T_c^l for $l = 1, \cdots, L$ is assumed to be the same as T_c. T_c is assumed to be known at the receiver, considered as prior knowledge in the system. Let $\zeta_l = \{i | \mathbf{x}_{cd}[u] = \mathbf{x}_p[l]\}$ represent the set of indices of coded data vectors which equal to the pilot vector $\mathbf{x}_p[l]$. Thus, the received vectors with indices in ζ_l can be written as the following form

$$\mathbf{y}[i] = \mathbf{H}\mathbf{x}_p[l] + \mathbf{w}[i], i \in \zeta_l \tag{2}$$

Let $\zeta_l^{\mathbf{y}} = \{\mathbf{y}[i] | \forall i \in \zeta_l\}$ represent the set of receive vectors that correspond to pilot vector $\mathbf{x}_p[l]$. Therefore, some received vectors in $\mathbf{Y}$ are assigned into subsets $\zeta_l^{\mathbf{y}}, l = 1, \cdots, L$, respectively. The optimal channel estimation for (2) based on maximum likelihood criteria is given by

$$\hat{\mathbf{H}} = \arg\min_{\mathbf{H}} \sum_{\mathbf{y} \in \zeta_l^{\mathbf{y}}} \|\mathbf{y} - \mathbf{H}\mathbf{x}_p[l]\|^2 \tag{3}$$

with $l = 1, \cdots, L$. Based on (3), we mainly focus on the clusters $\zeta_l^{\mathbf{y}}, l = 1, \cdots, L$, which correspond to the set of indices of pilot vectors $\mathbf{x}_p[l], l = 1, \cdots, L$.

In KMC [12], each observation is assigned to its closest centroid by calculating and comparing the Euclidean distances, which means that every observation in each cluster is densely distributed around its centroid in terms of Euclidean distances. The proposed E-KMC algorithm operates as follows. First, L centroids are initialized by the L selected columns of $\mathbf{Y}_P$. For the size of each cluster denoted as T_c, it is feasible that each centroid is surrounded by T_c receive vectors. Second, after the initialization of centroids, the l-th cluster $\zeta_l^{\mathbf{y}}$ can be determined by grouping the T_c nearest observations to l-th centroid. In E-KMC, the cluster $\zeta_l^{\mathbf{y}}$ is obtained with the aid of its cluster size and centroid. The observations with the same channel and data vectors are grouped into a cluster, which is capable of achieving a more accurate channel estimate by averaging in the following. Third, the average of the observations in each cluster is calculated.

Finally, the channel matrix in (3) can be calculated by

$$\hat{\mathbf{H}} = \bar{\mathbf{y}}_\zeta \mathbf{X}_P^H \left(\mathbf{X}_P \mathbf{X}_P^H\right)^{-1} \tag{4}$$

where $\bar{\mathbf{y}}_\zeta = [\bar{\mathbf{y}}_1, \cdots, \bar{\mathbf{y}}_l, \cdots, \bar{\mathbf{y}}_L]$ is an $N \times L$ matrix, with $\bar{\mathbf{y}}_l$ denoting the average of the observations in cluster $\zeta_l^{\mathbf{y}}$. Thanks to the clustering and the average operation, the noise power in $\bar{\mathbf{y}}_\zeta$ is much lower than that of the original centroids, which suggests that a more accurate channel estimate can be obtained. The proposed E-KMC algorithm is depicted in Algorithm 1.

Algorithm 1 : E-KMC based channel estimator

Input:
The pilot matrix $\mathbf{X}_P$;
The signal matrix $\mathbf{Y}$ received at CPU;
The size of each cluster T_c;
Output:
The estimated channel $\hat{\mathbf{H}}$;
Choose L columns of $\mathbf{Y}_P$ as centroids of L clusters, which are denoted as $\mathbf{y}[l], l = 1, \cdots, L$. The L clusters are denoted as $\zeta_l^{\mathbf{y}}, l = 1, \cdots, L$;
for $l = 1, \cdots, L$ **do**
  Calculate the Euclidean distances between the received signals and the centroid of each cluster: $D[j] = \|\mathbf{y}[j] - \mathbf{y}[l]\|^2, j = 1, \cdots, T$;
  Define the set of the closet T_c received vectors as the cluster $\zeta_l^{\mathbf{y}}$;
end for
Calculate the average of each cluster $\bar{\mathbf{y}}_l, l = 1, \cdots, L$, and $\bar{\mathbf{y}}_\zeta = [\bar{\mathbf{y}}_1, \cdots, \bar{\mathbf{y}}_l, \cdots, \bar{\mathbf{y}}_L]$ is obtained.
Compute the estimated channel by $\hat{\mathbf{H}} = \bar{\mathbf{y}}_\zeta \mathbf{X}_P^H \left(\mathbf{X}_P \mathbf{X}_P^H\right)^{-1}$;
return $\hat{\mathbf{H}}$;

It is noteworthy that the proposed E-KMC algorithm forms only L clusters of size T_c (the minimum value of L is K as stated in Subsection III-B), in contrast to M^K clusters of dynamic cluster size by KMC [10] [11]. Furthermore, as clustering requires the number of observations (frame length in this use case) to be much larger than the total number of clusters, E-KMC is applicable to a system with a relatively short frame length to meet low latency demands, while KMC is not. Hence, the proposed E-KMC algorithm is suitable for a wide range of CF massive MIMO system with flexible requirements on the number of transmit antennas, modulation order and frame length.

B. Pilot Design

Pilots are carefully designed in this subsection to achieve an accurate channel estimate but with much lower training overhead and computational complexity. To achieve both high spectral efficiency and high estimation accuracy, the length of the pilot sequence L is set to $L = K$, which is the shortest length allowable, since there are K unknown vectors from K sensors in the system needed to be solved by at least K equations through linear approaches.

To reduce the computational complexity of equation (4), it is efficient to design the pilot symbols carefully such that its pseudo inverse defined as $\mathbf{X}_P^{\dagger} = \mathbf{X}_P^H\big(\mathbf{X}_P\mathbf{X}_P^H\big)^{-1}$ is the sparsest. To guarantee a unique and low-complexity solution, the number of elements in the pseudo inverse matrix $\mathbf{X}_P^{\dagger}$ being zero or close to zero should be

$$Z = K(K-2) \tag{5}$$

Moreover, since there is an inverse operation in (4), the noise in $\bar{\mathbf{y}}_{\zeta}$ might be enhanced and propagated into the following channel estimate. To avoid the noise enhancement, we carefully design the pilots to minimize the power of $\mathbf{X}_P^{\dagger}$ which is defined as $\Theta_e = \mathrm{tr}\left\{\mathbf{X}_P^{\dagger}{\mathbf{X}_P^{\dagger}}^H\right\}$.

In addition, as mentioned above, K clusters are obtained by calculating the Euclidean distances between the received signal and the centroids of K clusters. Since the columns of received pilot symbols $\mathbf{y}[l], l = 1, \cdots, K$ are selected to be the centroids of K clusters, the performance of clustering can be improved by increasing the distances between pilot vectors $\mathbf{x}_p[l], l = 1, \cdots, K$. The sum of the distances between any two pilot vectors is defined as

$$\Theta_d = \sum_{m=1}^{K}\sum_{\substack{n=1\\n\neq m}}^{K}\left\|\mathbf{x}[m]-\mathbf{x}[n]\right\|^2 \tag{6}$$

Hence, the pilot design can be considered as an optimization problem which can be formulated as

$$\begin{aligned} &P1: \tilde{\mathbf{X}}_P = \arg\min_{\mathbf{X}_P\in\mathbb{C}} \alpha\Theta_e - \eta\Theta_d \\ &\text{Subject to (5)} \end{aligned} \tag{7}$$

where $\tilde{\mathbf{X}}_P$ denotes the optimal designed pilot matrix, α and η are multi-objective weights utilized to achieve a trade-off between the energy of $\mathbf{X}_P^{\dagger}$ and the sum of the distances between any two pilot vectors, and the pilot matrix should be full rank here to ensure (4) has a valid solution. However, it is extremely difficult to find the optimal solution to problem (7) directly. Thus we split the original optimization problem $P1$ into two simple subproblems $P1.1$ and $P1.2$, and then find a suboptimal solution to (7). Furthermore, it is noteworthy that the optimization of noise suppression is more important than the minimization of distance since the noise suppression affects both noise power and the distances, which suggests that $\alpha \gg \eta$. Thus, the optimization for the noise suppression subject to the lowest complexity is solved first which can be written as

$$\begin{aligned} &P1.1: \bar{\mathbf{X}}_P = \arg\min_{\mathbf{X}_P\in\mathbb{C}} \mathrm{tr}\left\{\mathbf{X}_P^{\dagger}{\mathbf{X}_P^{\dagger}}^H\right\} \\ &\text{Subject to (5)} \end{aligned} \tag{8}$$

Then, based on the solution set to P1.1, the subproblem about the distances among pilot vectors is formulated by

$$P1.2: \breve{\mathbf{X}}_P = \arg\max_{\bar{\mathbf{X}}_P\in\mathbb{C}_{\bar{\mathbf{X}}_P}} \sum_{m=1}^{K}\sum_{\substack{n=1\\n\neq m}}^{K}\left\|\bar{\mathbf{x}}[m]-\bar{\mathbf{x}}[n]\right\|^2 \tag{9}$$

where $\breve{\mathbf{X}}_P$ denotes the suboptimal designed pilot matrix for equation (7), $\mathbb{C}_{\bar{\mathbf{X}}_P}$ represents the solution set to (8), and $\bar{\mathbf{X}}_P$ is defined as $\bar{\mathbf{X}}_P = [\bar{\mathbf{x}}[1], \cdots, \bar{\mathbf{x}}[K]]$. Multiple suboptimal solutions to problem (7) might exist when the same distances are shared among pilot vectors.

We use $K = 6$ as an example in the following and assume that $\{c_1, c_2, \cdots, c_{11}\} \in \mathbb{C}$, where $\mathbb{C}$ is the constellation set. We describe how we determine the suboptimal pilot for M-PSK or M-QAM in the following.

First, Z is equal to 24 as stated in (5), so that each channel vector should be easily obtained by only two columns of $\bar{\mathbf{y}}_{\zeta}$ through (4). One possible pilot matrix set that meets the constraint $Z = 24$ can be depicted as

$$\breve{\mathbf{X}}_{\text{Pe}} = \begin{bmatrix} c_1 & c_2 & c_4 & c_6 & c_8 & c_{10} \\ c_1 & c_2 & c_4 & c_6 & c_8 & c_{11} \\ c_1 & c_2 & c_4 & c_6 & c_9 & c_{11} \\ c_1 & c_2 & c_4 & c_7 & c_9 & c_{11} \\ c_1 & c_2 & c_5 & c_7 & c_9 & c_{11} \\ c_1 & c_3 & c_5 & c_7 & c_9 & c_{11} \end{bmatrix} \tag{10}$$

Then, considering the property of the pseudo-inverse of $\mathbf{X}_{\text{Pe}}$, to ensure that $\mathbf{X}_{\text{Pe}}^{\dagger}$ has the minimum power, c_i with $i = 1, 2, \cdots, 11$ is chosen to be the symbol with maximum power in $\mathbb{C}$ and the distance between c_j and c_{j+1} with $j = 2, 4, 6, 8, 10$ should be the largest. Two elements in the set $\{c_1, c_2, \cdots, c_{11}\}$ might be equal if the number of maximum-power signals is smaller than 11. Finally, the distances among columns in $\breve{\mathbf{X}}_{\text{Pe}}$ are set to be the largest in order to greatly enhance the clustering performance through search.

As can be seen in the following sections, $\breve{\mathbf{X}}_{\text{Pe}}$ is a near-optimal solution to (7), which not only achieves a superior MSE performance of channel but also approaches the theoretical lower bound resented in Section IV, while with a lowest computational complexity, in comparison to the existing methods in [6], [9] and [11]. Moreover, since the pilot vectors are utilized to determine centroids in E-KMC algorithm, ambiguities are avoided.

IV. Performance and Complexity Analysis

A. Lower Bound on MSE of Channel Estimation

The normalized MSE of channel estimation is defined as

$$\text{MSE}_{\text{normalized}} = \frac{1}{KN}\mathbb{E}\left\{\left\|\mathbf{H}-\hat{\mathbf{H}}\right\|^2\right\} \tag{11}$$

The theoretical lower bound of normalized MSE of the proposed semi-blind E-KMC based channel estimation is given by

$$\Delta = \frac{N_0 K}{T_c \varpi} \tag{12}$$

where $\varpi = \mathrm{tr}\left\{\mathbf{X}_P{\mathbf{X}_P}^H\right\}$ is considered as the total power of pilot matrix [15].

Due to space limitation, the proof of (12) is not included in this paper.

TABLE I

COMPLEXITY ANALYSIS (K: NUMBER OF SENSORS, N: TOTAL NUMBER OF APS, M: MODULATION ORDER, L: PILOT LENGTH, T_{KMC}: FRAME LENGTH REQUIRED IN [11], T: FRAME LENGTH FOR E-KMC AND OTHERS [6] [9], T_c: NUMBER OF RECEIVED VECTORS OF EACH CLUSTER, R_{EM}: NUMBER OF ITERATIONS REQUIRED IN [9], R_{KMC}: NUMBER OF ITERATIONS REQUIRED IN [11]. EST.: ESTIMATION, DET.: DETECTION)

Item	E-KMC + MMSE	EM [9] + MMSE	ICA [6]	KMC [11]
Channel est.	$O\left(NKT + KT_c\right)$	$O\left(N^2\left(T - L + KR_{\text{EM}}\right)\right)$	$O(N^2K + NKT$ $+TK^2 + TK + TK^2)$	$O(T_{KMC}M^K R_{\text{KMC}})$
Signal det.	$O\left(K^3\right)$	$O\left(K^3\right)$		
Total	$O(NKT + KT_c + K^3)$	$O(N^2\left(T - L + KR_{\text{EM}}\right)$ $+K^3)$	$O(N^2K + NKT$ $+TK^2 + TK + TK^2)$	$O(T_{KMC}M^K R_{\text{KMC}})$

TABLE II

NORMALIZED NUMERICAL COMPLEXITY ($K = 6$, $N = 64$, $M = 4$, $L = 6$, $T_{KMC} = 81920$, $T = 1000$, $T_c = 50$, $R_{\text{EM}} = 5$, $R_{\text{KMC}} = 1000$. EST.: ESTIMATION, DET.:DETECTION)

Item	E-KMC + MMSE	EM [9] + MMSE	ICA [6]	KMC [11]
Total	1	10.6	1.2	8.3×10^5

B. Complexity Analysis

In Table I, the computational complexity of the proposed E-KMC scheme in comparison to existing methods [6] [9] [11] is presented. The complexity order of clustering operation and pseudo-inverse operation of pilots (4) in E-KMC is $O\left(NKT + KT_c\right)$ and $O(NK)$, respectively. Thus, the computational complexity of the proposed scheme is $O\left(NKT + KT_c\right)$. For a fair comparison, all the above schemes are assumed to share a common pilot. The pilot design requires a complexity of $O(M^K)$ to solve (7), which is much less than the complexity of $O(M^{K\times K})$ by exhaustive search. As the pilot is designed offline, it is not counted in the analyzed total complexity.

Normalized numerical complexity is shown in Table II, with $K = 6$ sensors, $N = 64$ APs, $M = 4$ (QPSK modulation), $L = 6$ pilot symbols, and frame length $T = (M^K) \times 20 = 81920$ for KMC [11] and $T = 1000$ for E-KMC and others [6] [9]. Each cluster of the proposed E-KMC estimator has $T_c = 50$ elements. The maximum number of iterations allowable for EM [9] and KMC [11] are set to $R_{\text{EM}} = 5$ and $R_{\text{KMC}} = 1000$, respectively. The proposed E-KMC scheme is more computationally efficient than the approaches in [6], [9] and [11]. In particular, it achieves nearly a 10^6-times complexity reduction over KMC [11]. This is because that E-KMC requires only one iteration to assign each cluster the group of received signal vectors nearest to the centroid, where the means are calculated only once with the aid of a short pilot. While KMC [11] has to conduct cluster assignment and means update iteratively in the 4096-cluster regime.

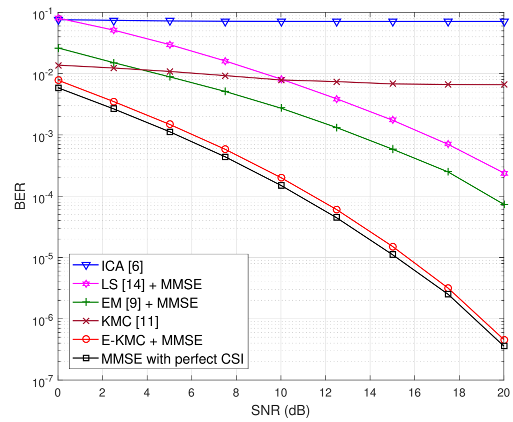

Fig. 2. BER of the proposed E-KMC based semi-blind channel estimation, with $K = 6$ sensors, $N = 64$ APs and QPSK modulation.

V. SIMULATION RESULTS

We use Monte-Carlo simulations to demonstrate the effectiveness of the proposed E-KMC scheme. The APs and sensors are uniformly distributed within an area of 500×500 m^2. The shadowing model and the three-slope path loss model [1] for large-scale fading are adopted. The same setup as for Table II is used except that 16-QAM modulation ($M = 16$) is used for Fig. 3. It is noteworthy that the required frame length for KMC [11] is $T = 81920$, which is approximately 82-fold longer than that of E-KMC with $T = 1000$ only. The designed pilot matrix in (10) is utilized in E-KMC. The ambiguity of KMC is mitigated with ideal CSI as assumed in [11]. The SNR is defined as the average of ratio of the received signal power to noise power at APs.

Fig. 2 demonstrates the BER performance of the proposed efficient E-KMC based semi-blind channel estimation method with QPSK modulation. It provides a BER performance that is much better than that of ICA [6], EM [9], KMC [11] and LS [14] methods, and close to the case with perfect CSI. The proposed E-KMC algorithm requires only 6 clusters and one iteration, while the KMC approach [11] requires 4096 clusters and does not converge even after 1000 iterations.

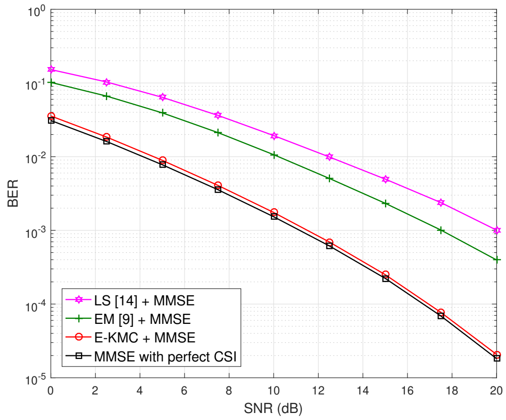

Fig. 3. BER of the proposed E-KMC based semi-blind channel estimation, with $K = 6$ sensors, $N = 64$ APs and 16-QAM modulation.

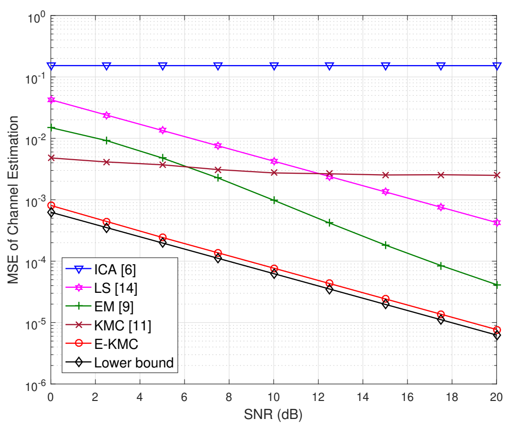

Fig. 4. Normalized MSE of channel estimation of the proposed E-KMC based semi-blind channel estimation, with $K = 6$ sensors, $N = 64$ APs and QPSK modulation.

ICA [6] suffers error floor due to insufficient ambiguity elimination.

Fig. 3 shows the BER performance with 16-QAM modulation, where similar trends to Fig. 2 can be observed. The complexity of KMC [11] in this case is prohibitive, with $16^6 = 1.7 \times 10^7$ clusters required, and therefore is not shown in Fig. 3. The performance of ICA [6] is also not shown as its ambiguity elimination method is for BPSK and QPSK only.

Fig. 4 illustrates the normalized MSE performance with QPSK modulation. With the pilot design presented in Subsection III-B, the lower bound on MSE derived in (12) is very close to the simulation result. The proposed E-KMC scheme demonstrates a much higher channel estimation accuracy than the ICA [6], EM [9], KMC [11] and LS [14] approaches, with a training overhead of only 0.6%.

VI. Conclusion

A low-complexity E-KMC algorithm has been proposed for semi-blind channel estimation at CPU of an uplink CF massive MIMO system in factory automation. The E-KMC algorithm achieves significant complexity reduction over the KMC algorithm [11], which is approximately 10^6-fold in a 6×64 QPSK system as the latter algorithm does not converge even after 1000 iterations. The BER performance of the E-KMC based receiver is very close to the perfect CSI case, and much better than those of the approaches based on ICA [6], EM [9], KMC [11] and LS [14]. A short pilot with very low training overhead is designed to find the centroids in clustering with dramatic complexity reduction over exhaustive search, while achieving a near-optimal performance. In conclusion, the proposed E-KMC based semi-blind scheme is applicable to an uplink CF massive MIMO system for factory automation with a wide range of the number of sensors, the number of APs, modulation order and frame length.

References

[1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834-1850, May 2017.

[2] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99878-99888, Jul. 2019.

[3] G. Hampel, C. Li and J. Li, "5G ultra-reliable low-latency communications in factory automation leveraging licensed and unlicensed bands," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 117-123, May 2019.

[4] D. Maryopi, M. Bashar and A. Burr, "On the uplink throughput of zero forcing in cell-free massive MIMO with coarse quantization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 7, pp. 7220-7224, Jul. 2019.

[5] E. Bjornson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, to be published.

[6] L. Shen, Y. Yao, H. Wang and H. Wang, "ICA based semi-blind decoding method for a multicell multiuser massive MIMO uplink system in Rician/Rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7501-7511, Nov. 2017.

[7] A. S. Guerreiro and G. Fraidenraich, "Phase corrected ICA for uplink massive MIMO," *IEEE Sig. Process. Lett.*, vol. 25, no. 12, pp. 1810-1814, Dec. 2018.

[8] Y. He et al., "Deep-reinforcement-learning-based optimization for cache-enabled opportunistic interference alignment wireless networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10433-10445, Nov. 2017.

[9] E. Nayebi and B. D. Rao, "Semi-blind channel estimation for multiuser massive MIMO systems," *IEEE Trans. Sig. Process.*, vol. 66, no. 2, pp. 540-553, Jan. 2018.

[10] H. Liang, W. Chung and S. Kuo, "Coding-aided K-Means clustering blind transceiver for space shift keying MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 103-115, Jan. 2016.

[11] L. Zhang and M. Jin, "A constrained clustering-based blind detector for spatial modulation," *IEEE Commun. Lett.*, vol. 23, no. 7, pp. 1170-1173, Jul. 2019.

[12] X. Wu et al., "Top 10 algorithms in data mining," *Knowl. & Inf. Systems*, vol. 14, no. 1, pp. 1-37, Jan. 2008.

[13] Rui Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Net.*, vol. 16, no. 3, pp. 645-678, May 2005.

[14] S. M. Kay, *Fundamentals of statistical signal processing: Estimation theory*, Upper Saddle River, NJ: Prentice-Hall, 1993.

[15] M. Biguesh and A. B. Gershman, "Training-based MIMO channel estimation: a study of estimator tradeoffs and optimal training signals," *IEEE Trans. Sig. Process.*, vol. 54, no. 3, pp. 884-893, Mar. 2006.