SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

December 2023, Vol. 66 222304:1–222304:13
https://doi.org/10.1007/s11432-023-3822-9

Robust channel estimation based on the maximum entropy principle

Zhengyang HU, Jiang XUE*, Feng LI, Qian ZHAO, Deyu MENG & Zongben XU

School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049*, China*

Received 10 April 2023/Revised 15 June 2023/Accepted 29 June 2023/Published online 1 November 2023

Abstract Channel estimation (CE) is one of the crucial and fundamental elements of signal processing, especially considering the requirement of high accuracy in future wireless communication systems. Most traditional CE algorithms are explored under the assumption of Gaussian white noise, which limits the algorithms performance in real wireless communication situations. In this work, a novel self-adaptive CE algorithm based on the maximum entropy principle (MEP) was studied, which analyzes the statistical components of an arbitrary noise environment. In addition, an MEP channel-based signal estimation algorithm was studied. Furthermore, the statistical characteristics of channels were considered the regularization terms in the objective function for providing prior information and further increasing the accuracy. It was found that the proposed algorithm not only provides accurate CE but also reduces pilot consumption by using estimated signal data as pseudo pilots. The superior features of the proposed method concerning CE accuracy, pilot consumption, and robustness were confirmed through Monte Carlo simulations.

Keywords channel estimation, maximum entropy principle, pseudo pilot, channel statistic characteristic, colorful noise

Citation Hu Z Y, Xue J, Li F, et al. Robust channel estimation based on the maximum entropy principle. Sci China Inf Sci, 2023, 66(12): 222304, https://doi.org/10.1007/s11432-023-3822-9

1 Introduction

Channel estimation (CE) plays a critical role in the physical layer of fifth generation (5G) systems [1]. While the 5G systems' performance is substantially limited by propagation channels [2], the accurate channel state information (CSI) is paramount in a number of communication technologies, such as massive multiple input multiple output (MIMO) [3], precoding [4], and adaptive transmission [5]. Thus, CE has garnered enormous interest from researchers.

Conventional methods for CE assume that the noise obeys a Gaussian distribution, which is valid in a long time interval that the number of noise components is large enough to satisfy the center limit theorem. Numerous studies have been explored based on this assumption over the past decades. A research group [6] proposed an algorithm combining the l_1-regularized least square and the l_2 minimum mean square error CE techniques, which decreases the bit error rate (BER) compared with l_2 multiburst CE. While the spacing of the antenna array is small and very few paths can arrive at the base station, the channel covariance matrices have low-rank properties [7]. Based on the rank deficiency, Fang et al. [8] used the minimum mean square error (MMSE) to realize CE with few overheads for training. Machine learning techniques have also been applied to CE [9]. Previous work [10] put forward the LDAMP network for CE, which combines the approximate message passing with image denoising methods. Another work [11] introduced an end-to-end scheme for direction-of-arrival estimation and CE, improving the performance of CE as well as having low computation complexity. Investigations have also been performed on combining CE with other modules or information to improve performance [12, 13]. The aforementioned methods generally assume that the noise obeys a Gaussian distribution. However, the noise component may not be large enough to satisfy the center limit theorem in real world scenarios. In addition, CE can be improved by considering the interference and noise jointly in specific scenarios [14].

*Corresponding author (email: x.jiang@xjtu.edu.cn)

Several researchers have investigated nonGaussian noise in a communication environment to overcome the aforementioned problem. Previous work [15] modeled the probability density function of the additive noise as a finite mixture of Gaussian (MoG) distribution in the signal processing pipeline. Another work [16] suggested a nonparametric likelihood-based CE method with MoG noise. Another research group [17] considered the MoG noise in array processing problems and developed an algorithm for estimating the source locations, signal waveforms, and noise distribution parameters. All these algorithms are based on MoG due to its effectiveness in the approximation. However, the MoG approximation cannot decide the number of mixture terms adaptively. A proposed robust CSI estimation method [14] was capable of determining the mixture term adaptively with a penalized term. Nevertheless, the channel statistical information and signal information have not been fully utilized. In addition, the algorithm in [14] pre-sets the formulation of nonGaussian noise, and it may lead to significant bias from the practical noise distribution.

The maximum entropy principle (MEP) can be employed for solving this problem. Based on the MEP, without the complete information of a distribution, the distribution with the maximized entropy corresponds to the object distribution [18]. Instead of using the pre-set formulation of the noise distribution, this formulation can be approximated directly from data by using the MEP. A group of authors [19] utilized the MEP to estimate the unknown interference by designing a detector in code-division multiple access (CDMA) systems. A denoising method was designed by tightening the Cramer-Rao lower bound, which was attributed to the MEP [20].

Although MEP-based strategies for CE can realize accurate probability density function estimation, they are restricted by the pilot consumption. An intuitive idea to overcome the pilot limitation is leveraging data-aided methods to extract necessary information for different tasks. Wee et al. [21] used the detected symbols and pilot symbols for calculating the cross-correlation and reducing the estimation error, which is attributed to the time-varying channel. Some researchers [22] investigated a data-aided scheme for CE to explore both the pilot and decoded data. Ju et al. [23] applied a data-aided method to CE with pilot contamination by using the knowledge of large-scale coefficients among local cells.

Furthermore, the statistical information of an object can be considered in specific scenarios. For instance, MMSE is a widely used algorithm that leverages the channel statistical information in CE. By assuming that the angular spread of users is limited to a narrow region, a previous work [24] exploited the low-rank properties of channels in massive MIMO systems. The channel covariance matrices of any two users with nonoverlapped angular spread are asymptotically orthogonal to each other, and the pilot contamination can be reduced. With numerous antennas at the base station, the channel would exhibit sparsity in beamspace for massive MIMO systems. Accordingly, the sparsity of channels was enhanced in [25], such that the users with nonoverlapped angular spread could use the same training data to eliminate pilot contamination.

Considering the aforementioned issues of CE, a robust and accurate method using pilots is required. Therefore, an enhanced MEP (EMEP) algorithm is developed herein. Through modeling the environment probability density function using the maximum entropy model and employing the maximum likelihood estimation (MLE) method, a standard machine learning measurement can be established. For the statistical properties of Rayleigh fading channels, a l_2 regularized term in the derived objective function can be structured to encode the properties. Thus, the objective function is associated with information about the environment and channel statistical properties. In addition, two methods are proposed to optimize the derived objective function; however, these processes exploit only the pilot information. To leverage the nonpilot information of the received data, a rough channel estimator is used to help in estimating the nonpilot transmitted signal as a pseudo pilot. Pseudo pilot estimation is essentially a nonconvex optimization problem and cannot be solved efficiently. Thus, it was relaxed to a tractable and standard least absolute shrinkage selection operator (LASSO) problem [26]. Numerical results show that the proposed algorithm outperforms conventional least square (LS) and noise modeling-based MoG methods [14].

Compared with our previous work, this paper mainly contributes in the following aspects:

- The statistical properties of the Rayleigh fading channel are formulated as l_2 regularization and incorporated into the objective function, which considers the environmental information and channel statistical properties.

- Based on the proposed model, an EMEP algorithm is proposed to update the parameters in the model. In addition, a new method is proposed for the optimization of the derived objective function, which exhibits good performance for a large channel size (32×32).

- Nonpilot signals, which contain substantial information about the channel, are considered in EMEP to decrease the pilot consumption and increase CE accuracy.
- Through numerical simulations, the robustness of EMEP concerning different communication environments, channel sizes, lengths of the pilot sequence, and the estimation accuracy of parameters and the channel is verified, confirming the superiority of the proposed method.

The rest of this paper is organized as follows. In Section 2, the system model is described. The EMEP algorithm is discussed in Section 3. In Section 4, the simulation results of the proposed method are presented. Lastly, the conclusion of this paper is presented in Section 5.

Throughout this paper, the following notations are adopted. Let bold uppercase and bold lowercase letters denote matrices and vectors, respectively. $(\cdot)^{\mathrm{T}}$ denotes the transpose operator and $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose of a matrix or vector. $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix and $\mathbf{1}$ represents the matrix whose elements are all 1s. $\mathbb{C}^{n\times m}$ denotes the set of $n \times m$ complex matrices. $\mathcal{N}(\mu, \sigma^2)$ represents the Gaussian distribution with mean μ and variance σ^2. $\det(\cdot)$ denotes the determinant of a matrix. The notation A_{ij} returns the i, j-th element of the matrix $\boldsymbol{A}$ and $|\cdot|$ denotes the absolute value. $\mathbb{E}[\cdot]$ and $\ln(\cdot)$ denote the expectation operator and natural logarithm, respectively. $\triangleq$ is used to indicate the definition of the value and new variable.

2 System model

We consider a MIMO system with Rayleigh fading channels, N_r antennas at the receiver, and N_t antennas at the transmitter. In the training part, the length of the transmitted signal sequence is L and the length of the pilot sequence is l, where $L > l$. Therefore, we use $\vec{\boldsymbol{X}}_p \in \mathbb{C}^{N_t\times l}$ and $\vec{\boldsymbol{X}} \in \mathbb{C}^{N_t\times L}$ to denote the pilot signal matrix and transmitted signal matrix, respectively. The channel matrix can be presented by an $N_r \times N_t$ matrix $\vec{\boldsymbol{H}} \in \mathbb{C}^{N_r\times N_t}$. In this paper, we consider the complex communication environment with interference and colorful noise. The received signal at the receiver can be written as

$$\vec{\boldsymbol{Y}} = \vec{\boldsymbol{H}}\vec{\boldsymbol{X}} + \vec{\boldsymbol{E}}, \tag{1}$$

where $\vec{\boldsymbol{E}} \in \mathbb{C}^{N_r\times L}$ and $\vec{\boldsymbol{Y}} \in \mathbb{C}^{N_r\times L}$ are matrices of the communication environment and the received signal, respectively. In this paper, we regard the communication environment as noise (unless specifically stated). $\boldsymbol{Y}_p$ is the pilot components of the received signal $\vec{\boldsymbol{Y}}$ associated with $\boldsymbol{X}_p$ as well as $\boldsymbol{E}_p$.

Considering $\vec{\boldsymbol{Y}}$, $\vec{\boldsymbol{X}}$, $\vec{\boldsymbol{H}}$ are complex matrices and the requirement of machine learning (ML), Eq. (1) can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{E},$$

where $\boldsymbol{Y} = [\mathrm{Re}(\vec{\boldsymbol{Y}}), \mathrm{Im}(\vec{\boldsymbol{Y}})]$, $\boldsymbol{H} = [\mathrm{Re}(\vec{\boldsymbol{H}}), \mathrm{Im}(\vec{\boldsymbol{H}})]$, $\boldsymbol{E} = [\mathrm{Re}(\vec{\boldsymbol{E}}), \mathrm{Im}(\vec{\boldsymbol{E}})]$, and

$$\boldsymbol{X} = \begin{bmatrix} \mathrm{Re}(\vec{\boldsymbol{X}}) & \mathrm{Im}(\vec{\boldsymbol{X}}) \\ -\mathrm{Im}(\vec{\boldsymbol{X}}) & \mathrm{Re}(\vec{\boldsymbol{X}}) \end{bmatrix}$$

with $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ representing real and imaginary parts, respectively. Once we get $\boldsymbol{H}$, $\vec{\boldsymbol{H}}$ can be obtained easily. In addition, all the parameters in this paper are set to real numbers for convenience.

3 EMEP for channel estimation

In this section, we will introduce the MEP-based model for CE and the pseudo pilot estimation model, which are key components of EMEP, in Subsections 3.1 and 3.2, respectively. Then, the solutions for the above models and the EMEP algorithm are presented in Subsection 3.3.

3.1 MEP-based channel estimation model

According to the principle of maximum entropy [27], a standard measurement in the machine learning manner for CE has been derived in Theorem 1.

Theorem 1. $p(\epsilon)$ is the environment probability density function, $\hat{p}(\epsilon)$ is the empirical probability density function, $f_m \triangleq |\epsilon|^{p_m}$, and $p_m > 0, m = 1, \ldots, M$. Based on the MEP model [28] and MLE, the measurement can be derived as (6).

Proof. The model about $p(\epsilon)$ by MEP can be presented as

$$\begin{aligned} \min \quad & \int p(\epsilon)\ln p(\epsilon)\,\mathrm{d}\epsilon \\ \text{s.t.} \quad & \mathbb{E}_p[f_m] = \mathbb{E}_{\hat{p}}[f_m], \qquad m = 1,\ldots,M, \\ & \int p(\epsilon)\,\mathrm{d}\epsilon = 1, \end{aligned} \tag{2}$$

where $\mathbb{E}_p[f_m]$ and $\mathbb{E}_{\hat{p}}[f_m]$ are the the expectation of f_m, and the independent variable of f_m follows the distribution p and $\hat{p}$, respectively. According to the model (2), the maximum entropy distribution can be derived as

$$p_\lambda(\epsilon) = \exp\left(\lambda_0 - 1 - \sum_{m=1}^{M} \lambda_m f_m(\epsilon)\right). \tag{3}$$

The key point of derivation from (2) to (3) can be found in [28]. It is worth noticing that the constraints of (2) are equality constraints, so the sign difference of maximum entropy distribution in [28] and (3) does not change any problem. According to (3), we can derive the likelihood function with regard to ϵ as

$$l_\lambda(\epsilon) = \prod_{\mathcal{I}=0}^{N_r \times L} \exp\left(\lambda_0 - 1 - \sum_{m=1}^{M} \lambda_m f_m(\epsilon)\right). \tag{4}$$

Considering $\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}$, the logarithmic form of likelihood function can be derived as

$$l_\lambda(\boldsymbol{H}) = \sum_{i,j} \left(\lambda_0 - 1 - \sum_{m=1}^{M} \lambda_m f_m\left(Y_{ij} - (HX)_{ij}\right)\right), \tag{5}$$

which replaces terms related $\boldsymbol{E}$ as $\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}$, and the independent variable is transformed from $\boldsymbol{E}$ to $\boldsymbol{H}$. Then, we can drive the measurement[1)] as

$$l_\lambda(\boldsymbol{H}) = \sum_{m=1}^{M} \lambda_m \|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}\|_{p_m}^{p_m}. \tag{6}$$

Considering the robustness of the estimator and the pilot limitation, we employ regularization methods to improve the generalization ability of the estimator and the accuracy of a few pilots. From a model standpoint, the regularization is a way to achieve the structure risk minimization. In addition, it avoids the over-fitting problem with insufficient data and obeys Occam's razor principle [29]. From a Bayes estimation standpoint, the regularization corresponds to the prior. In this paper, we consider Rayleigh fading channels with components generated from independent and identical Gaussian distributions $\mathcal{N}(0,1)$ [30] for both real and imaginary parts. According to the maximum a posterior criterion [31], we can therefore encode the prior information of Rayleigh fading channels into a l_2 regularization term. Given the derived measurement (6), we can write the objective function as

$$l_\lambda(\boldsymbol{H}) = \sum_{m=1}^{M} \lambda_m \|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}\|_{p_m}^{p_m} + \rho\|\boldsymbol{H}\|_2^2, \tag{7}$$

where ρ is the weight of the prior information.

3.2 Pseudo pilot estimation

To improve the estimation accuracy and the spectrum resource efficiency, we present a data-aided estimation approach based on the MEP. By exploiting the non-pilot received signal, the channel estimated from the rough channel estimator can be used to approximate the non-pilot transmitted signal as pseudo pilots. The optimization problem for obtaining pseudo pilots can be formulated as

$$\arg\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}\|_2^2, \tag{8}$$

[1)] The p-norm of matrix is element-wise in this paper.

and the pseudo pilots can be calculated by LS as

$$\boldsymbol{X} = (\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{Y}. \tag{9}$$

However, LS ignores the constraint of $\boldsymbol{X}$ and decreases the accuracy of the solution due to the extended solution space. The accuracy of the estimated matrix $\boldsymbol{X}$ further impacts CE. Considering that the individual element of the transmitted signal is 1 or -1[2], we rewrite (8) as

$$\arg\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}\|_2^2 \quad \text{s.t.} \quad |X_{ij}| = 1. \tag{10}$$

However, the constraint of (10) is non-convex and this problem is not tractable. The optimization problem thus becomes a discrete optimization problem.

To retrieve a tractable optimization, $\boldsymbol{X}$ can be parameterized as $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}$, where

$$\boldsymbol{S} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & -1 & 1 & 0 & 0 \end{bmatrix}_{N_t \times 2N_t},$$

and

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} & \cdots & a_{nm} \end{bmatrix}_{2N_t \times N_s}.$$

After this transformation, the elements of the estimated transmitted matrix can be represented as $X_{ij} = S_{i,2i-1}A_{2i-1,j} + S_{i,2i}A_{2i,j}$. In the case that X_{ij} is correctly estimated, only one non-zero element exists in $\{A_{2i-1,j}, A_{2i,j}\}$, and $A_{2i-1,j} + A_{2i,j} = 1, \forall i = 1, 2, \dots, N_r, j = 1, 2, \dots, 2N_t$. Therefore, $\boldsymbol{1} = \boldsymbol{B}\boldsymbol{A}$, where

$$\boldsymbol{B} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 1 & 0 & 0 \end{bmatrix}_{N_r \times 2N_t}.$$

Given this transformation, matrix $\boldsymbol{A}$ is sparse. Therefore, the optimization problem can be written as

$$\arg\min_{\boldsymbol{A}} \|\bar{\boldsymbol{Y}} - \bar{\boldsymbol{H}}\boldsymbol{A}\|_2^2 + \gamma\|\boldsymbol{A}\|_1, \tag{11}$$

where

$$\bar{\boldsymbol{Y}} = \begin{bmatrix} \boldsymbol{Y} \\ \sqrt{\mu}\boldsymbol{1} \end{bmatrix}, \ \bar{\boldsymbol{H}} = \begin{bmatrix} \boldsymbol{H}\boldsymbol{S} \\ \sqrt{\mu}\boldsymbol{B} \end{bmatrix},$$

and $\mu > \max(\frac{1}{2}, \gamma)$ is a hyper-parameter. Then, $\boldsymbol{X}$ can be derived by $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}$. While this approximation retrieves non-integer estimations, we apply the hard decision operator to constrain the elements of the estimated matrix $X_{ij} = 1$ or -1, for $i = 1, 2, \dots, N_t, j = 1, 2, \dots, N_s$. The hard decision operator can be explained in the following form:

$$X_{ij} = \begin{cases} 1, & \text{if} \quad |X_{ij} - 1| \leqslant |X_{ij} + 1|, \\ -1, & \text{otherwise}. \end{cases} \tag{12}$$

[2] Here we only consider the BPSK modulation. In Appendix A, we explain that the proposed data estimation method can be used in a higher order modulation.

3.3 The EMEP algorithm

In this subsection, we present the details of the EMEP algorithm. Subsection 3.3.1 is the solving method corresponding to the MEP-based channel estimation model, i.e., the derived measurement (7) in Subsection 3.1. Subsection 3.3.2 is the solving method corresponding to the pseudo pilot estimation described in Subsection 3.2. Finally, the EMEP algorithm is summarized in Subsection 3.3.3.

3.3.1 *MEP-based estimator*

According to the objective function (6), we need to estimate $\boldsymbol{\lambda}$ and $\boldsymbol{H}$. The method for updating $\boldsymbol{\lambda}$ and the derivations in detail can be found in [32]. Given the estimated $\boldsymbol{\lambda}$, we can optimize the objective function to estimate $\boldsymbol{H}$. For straightforward understanding, we firstly consider the non-regularized part and rewrite the objective function as

$$l(\boldsymbol{H}) = \lambda_2\|\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}\|_2^2 + \sum_{\substack{m=1,\\ m\neq 2}}^{M}\lambda_m\|\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}\|_{p_m}^{p_m}. \tag{13}$$

We introduce $M-1$ slack variables $\boldsymbol{M}_m = \boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}$, and derive the problem by the augmented Lagrangian method as

$$\begin{aligned}\min_{\boldsymbol{H}} L = {} & \lambda_2\|\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}\|_2^2 + \frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m\left\|\boldsymbol{M}_m-\boldsymbol{Y}+\boldsymbol{H}\boldsymbol{X}+\frac{\boldsymbol{\Delta}_m}{\rho_m}\right\|_2^2 \\ & + \sum_{\substack{m=1,\\ m\neq 2}}^{M}\lambda_m\|\boldsymbol{M}_m\|_{p_m}^{p_m},\end{aligned} \tag{14}$$

where $\boldsymbol{\Delta}_m = \boldsymbol{M}_m-\boldsymbol{Y}+\boldsymbol{H}\boldsymbol{X}, m = 1,3,\ldots,M$. Due to the overwhelming number of variables in (14), we leverage the core idea of the alternating direction method of multipliers (ADMM) [33] to decompose the original optimization problem into subproblems by proposition 1.

Proposition 1. The problem (14) can be decomposed to subproblems (15) and (16).

Proof.

$$\begin{aligned}\min_{\boldsymbol{H}} L &= \lambda_2\|\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}\|_2^2 + \sum_{\substack{m=1,\\ m\neq 2}}^{M}\lambda_m\|\boldsymbol{M}\|_{p_i}^{p_i} + \frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m\|\boldsymbol{M}_m-\boldsymbol{Y}+\boldsymbol{H}\boldsymbol{X}+\frac{\boldsymbol{\Delta}_m}{\rho_m}\|_2^2 \\ &= \frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m\left(\sum_{ij}\left(M_{mij}-Y_{ij}+\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j+\frac{\delta_{mij}}{\rho_m}\right)^2\right)+\lambda_2\sum_{ij}\left(Y_{ij}-\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j\right)^2 \\ &= \frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m\left(\sum_{ij}2\left(M_{mij}+\frac{\delta_{mij}}{\rho_m}\right)\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j\right)+\left(\lambda_2+\frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m\right)\sum_{ij}\left(Y_{ij}-\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j\right)^2 \\ &= \frac{\frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m(\sum_{ij}2(M_{mij}+\frac{\delta_{mij}}{\rho_m})\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j)}{\lambda_2+\frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m}+\sum_{ij}\left(Y_{ij}-\boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}_j\right)^2 \\ &= \frac{\frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m(\sum_{ij}2(M_{mij}+\frac{\delta_{mij}}{\rho_m})\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j)}{\lambda_2+\frac{1}{2}\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m}+\sum_{ij}\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}_j\right)^2-2\sum_{ij}Y_{ij}\boldsymbol{h}^{\mathrm{T}}\boldsymbol{x}_j \\ &= \sum_{ij}2\left[\frac{\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m(M_{mij}+\frac{\delta_{mij}}{\rho_m})}{2\lambda_2+\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m}-Y_{ij}\right]\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j+\left[\frac{\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m(M_{mij}+\frac{\delta_{mij}}{\rho_m})}{2\lambda_2+\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m}-Y_{ij}\right]^2+\left(\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j\right)^2\end{aligned}$$

$$
\begin{aligned}
&= \sum_{ij}\left[\boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j - \left(Y_{ij} - \frac{\sum_{\substack{m=1,\\ m\neq 2}}^{M} \rho_m\left(M_{mij}+\frac{\delta_{mij}}{\rho_m}\right)}{2\lambda_2+\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m}\right)\right]^2 \\
&= \left\|\boldsymbol{H}\boldsymbol{X} - \left(\boldsymbol{Y} - \frac{\sum_{\substack{m=1,\\ m\neq 2}}^{M} \rho_m\left(\boldsymbol{M}_m+\frac{\boldsymbol{\Delta}_m}{\rho_m}\right)}{2\lambda_2+\sum_{\substack{m=1,\\ m\neq 2}}^{M}\rho_m}\right)\right\|_2^2 \\
&= \|\boldsymbol{\Delta}-\boldsymbol{H}\boldsymbol{X}\|_2^2. \qquad (15)
\end{aligned}
$$

Then, the optimization part for variable $\boldsymbol{M}_m$ can be presented as

$$\min_{\boldsymbol{M}_m}\frac{1}{2}\left\|\boldsymbol{M}_m - \left(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}-\frac{\boldsymbol{\Delta}_m}{\rho_m}\right)\right\|_2^2 + \frac{\lambda_m}{\rho_m}\|\boldsymbol{M}_m\|_{p_m}^{p_m}, \qquad (16)$$

where $m = 1, 3, \ldots, M$.

Considering the regularization, Eq. (15) can be rewritten as

$$\min_{\boldsymbol{H}}\sum_j \|\boldsymbol{\Delta}_j - \boldsymbol{H}\boldsymbol{X}_j\|_2^2 + \rho\|\boldsymbol{H}\|_2^2. \qquad (17)$$

Then, let its gradient about $\boldsymbol{H}$ be zero to estimate $\boldsymbol{H}$ as

$$\boldsymbol{H} = \left(\sum_j \boldsymbol{\Delta}_j\boldsymbol{X}_j^{\mathrm{T}}\right)\left(\sum_j \boldsymbol{X}_j\boldsymbol{X}_j^{\mathrm{T}} + \rho\boldsymbol{I}\right)^{-1}. \qquad (18)$$

For each subproblem in (16), we consider its element-level formulation. A single subproblem can be represented as

$$\min_{M_{mij}} f = \frac{1}{2}\|z_{mij}-M_{mij}\|^2 + \lambda\|M_{mij}\|^{p_m}, \qquad (19)$$

where $z_{mij} = Y_{ij} - \boldsymbol{h}_i^{\mathrm{T}}\boldsymbol{x}_j - \frac{\delta_{mij}}{\rho_m}$, $\lambda = \frac{\lambda_m}{\rho_m}$, $i = 1, 2, \ldots, N_t$, and $j = 1, 2, \ldots, L$.

When $0 < p_m \leqslant 1$,

$$M_{mij} = \begin{cases} 0, & \text{if } |z_{mij}| < \tau_{p_m}(\lambda), \\ \operatorname{sgn}(z_{mij})S_{p_m}(|z_{mij}|;\lambda), & \text{if } |z_{mij}| > \tau_{p_m}(\lambda), \end{cases} \qquad (20)$$

where $S_{p_m}(|z_{mij}|;\lambda)$ is an iterative operator: $M_{ij}^{k+1} = |z_{mij}| - \lambda p_m (M_{mij}^k)^{p_m-1}$ and $\tau_{p_m}(\lambda) = (2\lambda(1-p_m))^{\frac{1}{2-p_m}} + \lambda p_m(2\lambda(1-p_m))^{\frac{p_m-1}{2-p_m}}$.

When $p_m > 1$, the subproblems are convex problems; we calculate the first-order and second-order derivatives of (19),

$$f' = M_{mij} - Y_{ij} + \lambda p_m |M_{mij}|^{p-1}\operatorname{sgn}(M_{mij}), \qquad (21)$$

$$f'' = 1 + \lambda p_m(p_m-1)|M_{mij}|^{p_m-2}. \qquad (22)$$

Then we can use the Newton method to approach M_{mij} as

$$M_{mij}^{k+1} = M_{mij}^k - \frac{f'(M_{mij}^k)}{f''(M_{mij}^k)}. \qquad (23)$$

ρ_m can be updated through the gradient method as

$$\rho_m^{k+1} = \rho_m^k - \alpha_2\frac{\|\boldsymbol{\Delta}_m\|_2^2 + \lambda_m\|\boldsymbol{M}_m\|_{p_m}^{p_m}}{{\rho_m^k}^2}, \qquad (24)$$

where α_2 is the step length. The above optimization method is defined as 'EMEP-A'.

'EMEP-A' decomposes the optimization problem into subproblems, which can be calculated in parallel. However, 'EMEP-A' introduces the auxiliary variables such as $\boldsymbol{\Delta}_m, \boldsymbol{M}_m, \rho_m$; each of these auxiliary variables is updated with iterations, which may accumulate the estimation error. To reduce the error, we rewrite the objective function (6) as

$$l(\boldsymbol{H}) = \sum_{i,j} W_{ij} \left(Y_{ij} - \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{x}_j\right)^2, \tag{25}$$

where $W_{ij} = \sum_{m=1}^{M} \lambda_m |E_{ij}|^{p_m - 2}$ and $\boldsymbol{W} \in \mathbb{R}^{N_r \times L}$. Therefore, $\boldsymbol{H}$ can be updated by the following iteration formulations:

$$W_{ij}^{k+1} = \sum_{m=1}^{M} \lambda |E_{ij}^{k} + \beta|^{p_m - 2}, \tag{26}$$

$$\boldsymbol{h}_i^{k+1\,\mathrm{T}} = \frac{\boldsymbol{X} \mathrm{diag}(w_{i1}^{k+1}, w_{i2}^{k+1}, \ldots, w_{il}^{k+1}) \boldsymbol{y}_i}{\boldsymbol{X} \mathrm{diag}(w_{i1}^{k+1}, w_{i2}^{k+1}, \ldots, w_{il}^{k+1}) \boldsymbol{X}^{\mathrm{T}}}, \tag{27}$$

$$E_{ij}^{k+1} = Y_{ij} - \boldsymbol{h}_i^{k+1\,\mathrm{T}} \boldsymbol{x}_j, \tag{28}$$

where the details can be found in [32]. After this step, we consider the regularization term that solely impacts the updates of $\boldsymbol{H}$. With iterations, $\boldsymbol{H}$ can be updated as

$$\boldsymbol{h}_i^{k+1\,\mathrm{T}} = \frac{\boldsymbol{X} \mathrm{diag}(w_{i1}^{k+1}, w_{i2}^{k+1}, \ldots, w_{iL}^{k+1}) \boldsymbol{y}_i}{\boldsymbol{X} \mathrm{diag}(w_{i1}^{k+1}, w_{i2}^{k+1}, \ldots, w_{iL}^{k+1}) \boldsymbol{X}^{\mathrm{T}} + \rho \boldsymbol{1}}. \tag{29}$$

Let 'EMEP-B' denote the above optimization method. 'EMEP-B' only introduces one auxiliary variable $\boldsymbol{W}$ to employ the re-weighting method [34], which has less error accumulation by iterative update compared with 'EMEP-A' due to the fewer auxiliary variables. However, only using the gradient information of the second-order term limits its performance when the channel size is large.

3.3.2 *Pseudo pilot estimation*

According to the optimization problem (11), there are many methods to solve the LASSO problem. We can rewrite the matrix into column vectors and approach them, respectively. Considering parallel computing, we use the ADMM to solve the problem (10). We represent the problem (11) as

$$\min_{\boldsymbol{A}, \boldsymbol{Z}} \left\{ \|\bar{\boldsymbol{Y}} - \bar{\boldsymbol{H}} \boldsymbol{A}\|_2^2 + \gamma \|\boldsymbol{Z}\|_1 \right\} \quad \text{s.t.} \quad \boldsymbol{A} = \boldsymbol{Z}. \tag{30}$$

For approaching $\boldsymbol{A}$ and $\boldsymbol{Z}$, we replace the 1-norm of matrix with 1-norm of vector and calculate each column of $\boldsymbol{A}$ and $\boldsymbol{Z}$, respectively. The iteration formulations are

$$\boldsymbol{a}_i^{k+1} := \arg\min_{\boldsymbol{a}_i} \left\{ \|\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{H}} \boldsymbol{a}_i\|_2^2 + \frac{\rho}{2} \left\| \boldsymbol{a}_i - \boldsymbol{z}_i^k + \boldsymbol{v}_i{}^k \right\|_2^2 \right\}, \tag{31}$$

$$\boldsymbol{z}_i^{k+1} := \arg\min_{\boldsymbol{z}_i} \left\{ \gamma \|\boldsymbol{z}_i\|_1 + \frac{\rho}{2} \left\| \boldsymbol{a}_i^{k+1} - \boldsymbol{z}_i + \boldsymbol{v}_i{}^k \right\|_2^2 \right\}, \tag{32}$$

$$\boldsymbol{v}_i^{k+1} := \boldsymbol{v}_i^k + \boldsymbol{a}_i^{k+1} - \boldsymbol{z}_i^{k+1}. \tag{33}$$

3.3.3 *Summarizing the EMEP algorithm*

We first use the MEP-based estimator with the pilot signal to estimate the channel for the pseudo pilot estimation. Then, the MEP-based estimator is used again with the pilot and pseudo pilot signal to estimate the channel. The algorithm can be summarized as Algorithm 1.

3.4 Complexity

The variables for solving the MEP-based model are described in closed forms. The complexity for pseudo pilot estimation increases linearly with the length of the transmitted signal sequence. Therefore, we can evaluate that the complexity of the proposed methods is $\mathcal{O}(I_1(N_r L + N_t L + N_t^3) + L)$ and $\mathcal{O}(I_2(N_r N_t L + N_t^3 N_r) + L)$ for 'EMEP-A' and 'EMEP-B', respectively, where I_1 is the iteration number of 'EMEP-A', and I_2 is the iteration number of 'EMEP-B'.

Algorithm 1 Algorithm for EMEP

```
Require: received signal $\boldsymbol{Y}$, pilot sequence $\boldsymbol{X}_p$, pilot part of received signal $\boldsymbol{Y}_p$, feature set $\boldsymbol{p}$;
Ensure: channel $\boldsymbol{H}$, parameter $\boldsymbol{\lambda}$, feature set $\boldsymbol{p}$;
Initialization: $\{\boldsymbol{H}, \boldsymbol{\lambda}, \boldsymbol{p}, \boldsymbol{E}_p\}$;
while not converged do
    Update $\boldsymbol{\lambda}$;
    Update $\{\boldsymbol{H}, \boldsymbol{p}, \boldsymbol{E}_p\}$ by (18), (20) or (23), (24) with $\boldsymbol{X}_p$ and $\boldsymbol{Y}_p$ ('EMEP-A'), or update $\{\boldsymbol{H}, \boldsymbol{p}, \boldsymbol{E}_p\}$ by (26), (28), (29) with $\boldsymbol{X}_p$ and $\boldsymbol{Y}_p$ ('EMEP-B');
end while
Estimate pseudo pilot $\boldsymbol{X}$ by (31)–(33) with $\boldsymbol{Y}$ and $\boldsymbol{H}$;
Update $\boldsymbol{H} = \boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}$, and $\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}$;
while not converged do
    Update $\boldsymbol{\lambda}$;
    Update $\{\boldsymbol{H}, \boldsymbol{p}, \boldsymbol{E}\}$ by (18), (20) or (23), (24) ('EMEP-A'), or update $\{\boldsymbol{H}, \boldsymbol{p}, \boldsymbol{E}\}$ by (26), (28), (29) ('EMEP-B');
end while
```

4 Simulation results

In this section, the performance and the adaption ability of the proposed EMEP method with additive nonGaussian noise are verified via Monte Carlo simulations.

The LS estimator is

$$\boldsymbol{H}_{\mathrm{LS}} = \boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}. \tag{34}$$

The linear minimum mean square error (LMMSE) estimator is

$$\boldsymbol{H}_{\mathrm{LMMSE}} = \boldsymbol{Y}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \sigma_e^2\boldsymbol{I})^{-1}, \tag{35}$$

where σ_e^2 is the average power of noise. The above two methods and MoG method [30] will be used to provide performance references. The signal-to-noise ratio (SNR) is applied to present the noise level and is defined as

$$\mathrm{SNR} = \frac{\sigma_x^2}{\sigma_e^2}, \tag{36}$$

where σ_x^2 and σ_e^2 are the average power of transmitted signals and noise, respectively.

In the training phase of the uplink channel, the user transmits pilot sequences of length l. We choose the normalized mean square error (NMSE) to evaluate the accuracy of channel estimation, which is

$$\mathrm{NMSE} = \frac{\|\boldsymbol{H} - \hat{\boldsymbol{H}}\|_2^2}{\|\boldsymbol{H}\|_2^2}, \tag{37}$$

where $\boldsymbol{H}$ and $\hat{\boldsymbol{H}}$ are the real and estimated channels.

4.1 Communication environment settings

To simulate the complex communication environment, we follow the generating method in [30] and generate the environment by MoG distribution as

$$\mathbb{P}(\epsilon) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\epsilon \mid \mu_k, \sigma_k^2), \tag{38}$$

where K is the number of Gaussian distributions, $\mathcal{N}(\epsilon \mid \mu_k, \sigma_k^2)$ stands for the Gaussian distribution with mean μ_k and variance σ_k^2. $\pi_k \geqslant 0$, and $\sum_{k=1}^{K} \pi_k = 1$ is the mixing proportion. Since the Laplace noise is a common assumption in indoor and outdoor communications, submarine transmission, and ultra-wide bandwidth wireless communication [35], we also generate the environment with Laplace distribution to verify the robustness of the proposed method in terms of communication environment. The Laplace distribution is defined as

$$\mathbb{P}(\epsilon) = \mathcal{L}(\epsilon \mid \mu_l, \sigma_l) \tag{39}$$

with mean μ_l and variance σ_l.

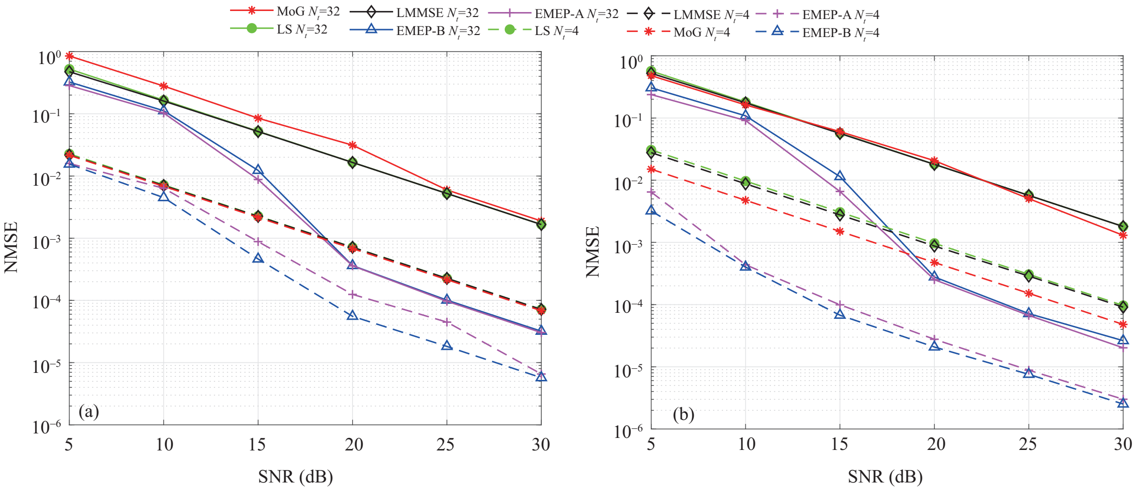

Figure 1 (Color online) NMSE versus SNR for the EMEP and several methods with (a) Case 1 and (b) Case 2.

4.2 Performance evaluation

In this subsection, we evaluate the NMSE performance of the EMEP algorithm in terms of different environment probability density functions, channel sizes, pilot lengths, and the number of transmitted signals through comparison experiments to study its robustness. Moreover, the symbol error rate (SER) performance is also presented to verify the effectiveness of EMEP, which is defined as

$$\mathrm{SER} = \frac{N_{se}}{N_s}, \tag{40}$$

where N_{se} and N_s denote the number of wrong symbols and the number of all symbols, respectively. Finally, the strong ability of EMEP in terms of distribution estimation is demonstrated through quantitative experiments.

We study the performance of EMEP with different environment probability density functions to investigate the robustness of using MEP to fit the environment. In addition, the differences of 'EMEP-A' and 'EMEP-B' are also studied. In Figure 1(a), the length of pilot sequences, the length of transmitted signal sequences, and the number of transmitted antennas and received antennas are set as $l = 50$, $L = 1000$, and $N_t = N_r \in \{4, 32\}$, respectively. The environment (Case 1) is generated as 45% is following Gaussian $\mathcal{N}(0, \sigma_1^2)$, 45% is following Gaussian $\mathcal{N}(2, \sigma_2^2)$, and the remaining 10% is following Gaussian $\mathcal{N}(-2, \sigma_2^2)$, where the corresponding SNR is in the range of $[5, 30]$ dB. In Figure 1, the environment (Case 2) is set as $\mathcal{L}(0, \sigma_l)$, the SNR varies in the range of [5, 30] dB, and other settings are as the same as the settings in Figure 1(a).

It is obvious that the NMSE increases for all methods when the number of antennas is increasing in Figures 1(a) and (b), due to the error accumulation caused by the increasing number of estimated parameters. However, EMEP (i.e., 'EMEP-A' and 'EMEP-B') outperforms the compared methods when the channel sizes are the same. Moreover, EMEP achieves the best performance among the compared methods with different environment probability density functions, which demonstrates its adaptive ability. We can also see that the performance gap between the EMEP-A and EMEP-B is narrowing with the increase of SNR when $N_t = 32$. Meanwhile, when $N_t = 4$, the performance gap is stable. In addition, from Figures 1(a) and (b), it can be seen that 'EMEP-A' achieves better performance when the channel size is 32×32, while 'EMEP-B' achieves better performance when the channel size is 4×4. This phenomenon may be attributed to that the estimation error accumulation caused by auxiliary variables affects significantly when the channel size is small, while only using gradient information of the second-order limits the performance when the number of estimated parameters, i.e., the channel size, is large. For simplicity, we set the small channel size, and 'EMEP' stands for using 'EMEP-B' to obtain channels in the rest of simulations.

We set the number of transmitted antennas, received antennas, and the length of transmitted signal sequences as $N_r = N_t = 4$ and $L = 1000$. The length of pilot sequences $l = 50$ and $l = 100$. The environment is generated as: 10% is following Gaussian $\mathcal{N}(0, \sigma^2)$, 20% is following Gaussian $\mathcal{N}(0, 0.01 \times$

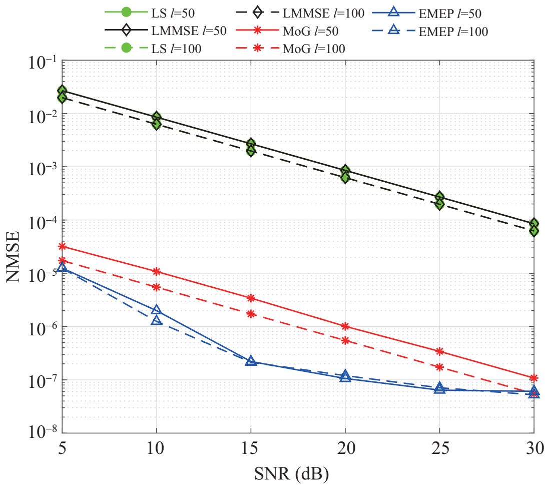

Figure 2 (Color online) NMSE versus SNR for different methods with different lengths of pilot sequences.

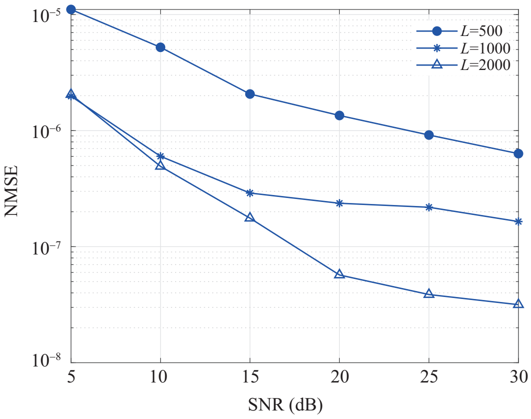

Figure 3 (Color online) NMSE versus SNR for EMEP with different L.

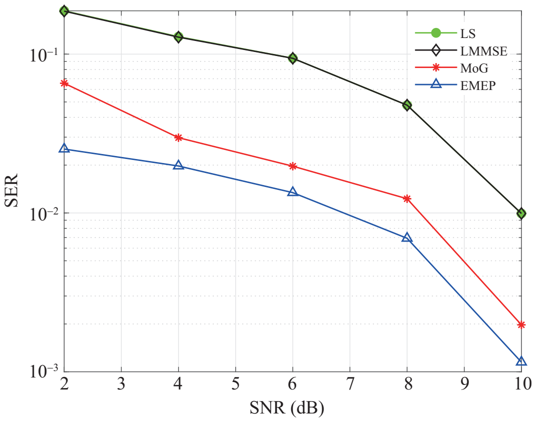

Figure 4 (Color online) SER versus SNR for different methods.

σ^2), and the remaining 70% is following Gaussian $\mathcal{N}(0, 0.0001 \times \sigma^2)$, where the corresponding SNR is in the range of $[5, 30]$ dB. In Figure 2, it is obvious that EMEP outperforms the compared methods with different lengths of pilot sequences l. It can also be seen that the performance increases with the increase in the number of lengths of pilot sequences for all methods, while the performances of EMEP with different lengths of pilot sequences have no significant gap. Due to the pseudo pilot estimation, EMEP can also utilize the estimated signal rather than only the pilot signal for CE, which means the performance of EMEP does not highly depend on the length of pilot sequences, and the good performance can be achieved by fewer numbers of lengths of pilot sequences.

In Figure 3, we compare the NMSE of the CE versus SNR for different lengths of transmitted signal sequences $L \in \{500, 1000, 2000\}$, and the pilot sequence length $l = 50$. Other settings are the same as the settings in Figure 2. From Figure 3, we can see that the NMSE decreases when the number of transmitted signals increases, especially when SNR is higher than 20 dB. It shows that increasing L can improve the upper limitation of EMEP. However, when the number of transmitted signals is large, the decrease of NMSE is not obvious at a low SNR regime. The reason for the mentioned phenomenon is that the pseudo pilot estimation is not accurate enough to gain the benefit of increasing the number of transmitted signals.

To further verify the effectiveness of the proposed method, we compare the SER of the proposed EMEP method with the compared methods in Figure 4. The lengths of transmitted signal sequences $L = 1000$, corresponding SNR is in the range of $[2, 10]$ dB, and other settings are the same as in Figure 3. Considering the fair comparison, we use the maximum likelihood detection method for all methods to calculate SER. The results in Figure 4 show that EMEP also outperforms all compared methods in terms

Table 1 Quantitative comparison of the true (denoted by "True") environment probability density functions and the estimated results (denoted by "Est.") in 3 different cases

	Type	$\boldsymbol{\lambda}$
Case 1	True	[1.414, 0]
	Est.	[1.4974, 4.99×10^{-5}]
Case 2	True	[0, 0.125]
	Est.	[0.0951, 0.116]
Case 3	True	[0, 0.5]
	Est.	[0.1591, 0.4836]

of SER, which further demonstrates that EMEP can obtain the accurate channel.

Table 1 shows the parameters estimation results of the true environment probability density functions. The true environment probability density functions of three cases are $\mathcal{L}(0, \frac{1}{\sqrt{2}})$, $\mathcal{N}(0, 4)$, and $\mathcal{N}(0, 1)$, respectively. The initialization of parameter $\boldsymbol{p}$ is all set as $[1, 2]$ in three cases. The estimated values of $\boldsymbol{\lambda}$ corresponding to the second order part and the first order part are almost zero in Cases 1 and 2, respectively, which shows that our algorithm extracts the different moment components of the noise distribution automatically. The parameters estimation error is around 10^{-1} in all three cases, which also shows the strong moment information extracting ability of our algorithm. The above results verify that the proposed noise modeling method can investigate the moment information of the noise and get insights into the environment.

5 Conclusion

In this work, the issues pertaining to CE in practical wireless communication systems were explored. Since complex environments and the required number of pilots limit the accuracy of CE, reducing pilot consumption and improving the robustness of CE are important. From the perspective of ML, a data-aided CE algorithm, namely EMEP, is proposed. The noise modeling method and MEP were exploited to approximate the probability density functions of the noise and derive the measurement for estimating channel, which can adaptively determine the moment component of noise in CE under complex communication environments. Further, the nonpilot signal in the CE structure was considered to reduce pilot consumption, and further increase the estimation accuracy. Notably, the proposed EMEP algorithm demonstrated robustness in different complex communication environments. Simulation results confirmed the superiority of EMEP concerning estimation accuracy, pilot consumption, and environment adaptation.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2020YFA0713900), Macao Science and Technology Development Fund (Grant No. 061/2020/A2), National Natural Science Foundation of China (Grant Nos. 61721002, 12226004, 62076196), and in part by Major Key Project of Peng Cheng Laboratory (Grant No. PCL2023AS1-2).

References

1 Yuan H, Kam P Y. Soft-decision-aided, smoothness-constrained channel estimation over time-varying fading channels with no channel model information. IEEE Trans Wireless Commun, 2017, 16: 73–86

2 Shafi M, Molisch A F, Smith P J, et al. 5G: a tutorial overview of standards, trials, challenges, deployment, and practice. IEEE J Sel Areas Commun, 2017, 35: 1201–1221

3 Larsson E G, Edfors O, Tufvesson F, et al. Massive MIMO for next generation wireless systems. IEEE Commun Mag, 2014, 52: 186–195

4 Sohrabi F, Liu Y F, Yu W. One-bit precoding and constellation range design for massive MIMO with QAM signaling. IEEE J Sel Top Signal Process, 2018, 12: 557–570

5 Zhang F, Sun S, Gao Q, et al. Enhanced CSI acquisition for FDD multi-user massive MIMO systems. IEEE Access, 2018, 6: 23034–23042

6 Takano Y, Juntti M, Matsumoto T. l_1 LS and l_2 MMSE-based hybrid channel estimation for intermittent wireless connections. IEEE Trans Wireless Commun, 2016, 15: 314–328

7 Xie H X, Gao F F, Jin S. An overview of low-rank channel estimation for massive MIMO systems. IEEE Access, 2016, 4: 7313–7321

8 Fang J, Li X, Li H, et al. Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems. IEEE Trans Wireless Commun, 2017, 16: 1935–1947

9 Qi C H, Dong P H, Ma W Y, et al. Acquisition of channel state information for mmWave massive MIMO: traditional and machine learning-based approaches. Sci China Inf Sci, 2021, 64: 181301

10 He H T, Wen C K, Jin S, et al. Deep learning-based channel estimation for beamspace mmWave massive MIMO systems. IEEE Wireless Commun Lett, 2018, 7: 852–855

11 Huang H, Yang J, Huang H, et al. Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system. IEEE Trans Veh Technol, 2018, 67: 8549–8560

12 Ma X, Gao Z, Gao F F, et al. Model-driven deep learning based channel estimation and feedback for millimeter-wave massive hybrid MIMO systems. IEEE J Sel Areas Commun, 2021, 39: 2388–2406

13 Xu W, Gao F F, Zhang J, et al. Deep learning based channel covariance matrix estimation with user location and scene images. IEEE Trans Commun, 2021, 69: 8145–8158

14 Zhang H P, Xue J, Meng D Y, et al. Robust CSI estimation under complex communication environment. In: Proceedings of IEEE International Conference on Communications, Shanghai, 2019. 1–6

15 Kozick R J, Blum R S, Sadler B M. Signal processing in non-Gaussian noise using mixture distributions and the EM algorithm. In: Proceedings of Asilomar Conference on Signals, Systems and Computers, Pacific Grove, 1997. 438–442

16 Bhatia V, Mulgrew B. Non-parametric likelihood based channel estimator for Gaussian mixture noise. Signal Processing, 2007, 87: 2569–2586

17 Sadler B M, Kozick R J. Maximum-likelihood array processing in non-Gaussian noise with Gaussian mixtures. IEEE Trans Signal Process, 2000, 48: 3520–3535

18 Jaynes E T. Information theory and statistical mechanics. Phys Rev, 1957, 108: 171–190

19 Saberali S M, Amindavar H. Nonlinear detector design for CDMA signals in the presence of unknown interferers using maximum entropy method and comparison with SIC. IEEE Commun Lett, 2014, 18: 737–740

20 Li H, Li X L, Anderson M, et al. A class of adaptive algorithms based on entropy estimation achieving CRLB for linear non-Gaussian filtering. IEEE Trans Signal Process, 2012, 60: 2049–2055

21 Wee J, Jeon W, Lee Y, et al. Pilot and data aided channel estimation for OFDM systems in rapidly time-varying channels. In: Proceedings of IEEE 69th Vehicular Technology Conference, Barcelona, 2009. 1–5

22 Ma J J, Ping L. Data-aided channel estimation in large antenna systems. IEEE Trans Signal Process, 2014, 62: 3111–3124

23 Ju M Y, Xu L, Jin L, et al. Data aided channel estimation for massive MIMO with pilot contamination. In: Proceedings of IEEE International Conference on Communications, Paris, 2017. 1–6

24 Yin H F, Gesbert D, Filippou M, et al. A coordinated approach to channel estimation in large-scale multiple-antenna systems. IEEE J Sel Areas Commun, 2013, 31: 264–273

25 Xie H, Gao F F, Zhang S, et al. A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model. IEEE Trans Veh Technol, 2017, 66: 3170–3184

26 Tibshirani R. Regression shrinkage and selection via the Lasso. J R Statistical Soc Ser B-Stat Methodol, 1996, 58: 267–288

27 Guiasu S, Shenitzer A. The principle of maximum entropy. Math Intell, 1985, 7: 42–48

28 Cover T M, Thomas J A. Elements of Information Theory. Hoboken: John Wiley & Sons, 1999

29 Myung I J, Pitt M A. Applying Occam's razor in modeling cognition: a Bayesian approach. Psychonomic Bull Rev, 1997, 4: 79–95

30 Du H, Deng Y, Xue J, et al. Robust online CSI estimation in a complex environment. IEEE Trans Wireless Commun, 2022, 21: 8322–8336

31 Bishop C M, Nasrabadi N M. Pattern Recognition and Machine Learning. Berlin: Springer, 2006

32 Hu Z Y, Xue J, Meng D Y, et al. MEP-based channel estimation under complex communication environment. In: Proceedings of IEEE International Conference on Communications, Dublin, 2020. 1–5

33 Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. FNT Machine Learn, 2010, 3: 1–122

34 Xu Y C, Liu X D, Shen Y L, et al. Multi-task learning with sample re-weighting for machine reading comprehension. 2018. ArXiv:1809.06963

35 Dang X, Huang Z, Zhu L, et al. Block Turbo code in white Laplacian noise. Acta Aeronaut Astronaut Sin, 2016, 37: 3494–3501

Appendix A Data estimation for higher order modulation

If one symbol contains k bits as $s_1, \ldots, s_k$, we change matrixes $\boldsymbol{B}$, $\boldsymbol{A}$, and $\boldsymbol{S}$ as

$$\boldsymbol{S} = \begin{bmatrix} s_1 & \ldots & s_k & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & 0 & s_1 & \ldots & s_k & 0 & \ldots & \ldots & 0 \\ & & & & & & \ddots & & & \\ 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & s_1 & \ldots & s_k \end{bmatrix}_{N_r \times kN_t},$$

$$\boldsymbol{B} = \begin{bmatrix} 1 & \ldots & 1 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 & 0 & \ldots & \ldots & 0 \\ & & & & & & \ddots & & & \\ 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & 1 & \ldots & 1 \end{bmatrix}_{N_r \times kN_t},$$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} & \cdots & a_{nm} \end{bmatrix}_{mN_t \times N_s}.$$

Other derivations are the same in Subsection 3.2.