SIMO Channel Estimation Using Space-Time Signal Subspace Projection and Soft Information

Shu Cai†, Tad Matsumoto‡, and Kehu Yang†

† ISN Lab, Xidian University
South Taibai Road 2, 710071, Xi'an, China
phone & fax: + (86) 29-8820-4424, email: {yang001, caishu}@xidian.edu.cn

‡ Centre for Wireless Communications, University of Oulu
FIN-90014, Oulu, Finland, and
Japan Advanced Institute of Science and Technology
1-1 Asahidai Nomi Ishikawa 923-1292, Japan
Email: †{caishu, yang001}@xidian.edu.cn, ‡ tadashi.matsumoto@ee.oulu.fi

***Abstract*— We consider the channel estimation of a time-slotted wireless communication system with a mobile user and a base station, where the base station employs an M-element ($M > 1$) antenna array. The uplink single-input multiple-output (SIMO) channel is usually estimated by training sequence within each time slot. To improve the estimation performance, the channel estimate is often refined by projecting it to the corresponding spatial signal subspace. However, this projection will not work when the number of resolvable multipath rays is larger than that of the antenna array elements, which makes the channel matrix full row rank. In this paper, we formulate the channel estimation under the space-time signal model for this full-row-rank case, and propose a new method by space-time signal subspace projection using both training and unknown data sequences. To further improve the accuracy of the channel estimate, the soft information fed back from the decoder can be used. By involving this soft information, we propose another new channel estimation method. This method approximately follows the maximum likelihood (ML) criterion and is therefore referred to as the approximated ML channel estimation. Numerical results show that these methods can be performed separately or jointly to improve the performance of channel estimation by training sequences.**

***Index Terms*—Channel estimation, maximum likelihood, signal subspace projection, soft information, semi-definite relaxation**

I. Introduction

It is known that packet based transmission is often used in time-slotted wireless communications systems for IP-based data services. In order to satisfy the specific requirements of each user's quality-of-service (QoS), such as the data rate, the latency, and the bit error rate, an antenna array at each base station is often employed. However, the associated uplink wireless channel always varies in space and time along with either movements of users or changes of surroundings, or both. In this case, the QoS will depend on the performance of channel estimation. It is obvious that higher QoS will require higher accuracy of channel estimates.

With the use of the training sequence (or midamble codes) within each time slot, the estimation of the uplink SIMO channel in the above systems can be straightforwardly performed under the least square (LS) criterion. This leads to the LS channel estimation. Notice that the SIMO channel may have specific properties in some situations, which will be helpful to improve the estimation performance. For instance, when the number of resolvable multipath rays is smaller than either that of the antenna array elements or the maximum delay of the channel in symbol periods, the uplink SIMO channel matrix becomes rank deficient. This rank deficiency is referred to as the low-rank property. The LS channel estimation without involving this low-rank property is called the unconstrained LS (ULS) [1].

By employing the low-rank property of the SIMO channel (see the analysis in [2], [3] and the references therein), the reduced rank (RR) channel estimation methods [1], [3]–[5] are proposed to improve the performance of the channel estimation with reduced number of unknown parameters. In the situations where angles and delays of multipath rays are slowly varying, the training sequences of multiple slots can be used to estimate these channel features by the multislot (MS) channel estimation methods [5]–[7]. As shown in [5], [6], the RR and MS methods are equivalent to the projection of the whitened unconstrained LS channel estimate onto the spatial or temporal signal subspace without employing the received signal corresponding to the unknown data sequence within each slot. To improve the performance of the spatial signal subspace estimation and so as the channel estimation, the received signal corresponding to the unknown data sequence are employed by the signal subspace projection (SSP) method [4]. However, these projection-based methods will not work when the number of resolvable multipath rays is larger than that of the antenna array elements, which makes the channel matrix full row rank.

In this paper, we formulate the channel estimation under the space-time signal model for this full-row-rank case, and propose a new method by space-time signal subspace projection, where the received signal corresponding to both training and unknown data sequences are employed to estimate the signal subspace. Moreover, we show that the new method based on

space-time signal subspace projection is intrinsically related to the maximum likelihood (ML) estimation.

We next consider channel estimation by using soft information. As we know, the turbo principle [21] related to the well-known turbo code decoding algorithm has been widely employed in practical applications such as turbo equalization and turbo channel estimation [9]–[12], where the soft information fed back from the decoder(s) are used as the additional information to enhance the performance. With the use of the soft information, the turbo equalization techniques [8] achieve the near-optimal performance with lower complexity, while the channel estimation methods, called the soft-based channel estimation methods, provide significant performance improvements over the receivers with either a single antenna [9]–[11] or an antenna array [12], [13]. Specifically, in [13] the MS method has been extended to exploit soft-valued data for channel estimation. Although this extension has the advantage of simple implementation [see (3.35) in [13]], it does not consider the impact of the additional noise on the estimation performance which arises from the data estimation errors. Based on the same signal model for the soft-valued data mentioned there and assuming that this additional noise is Gaussian distributed, we here consider the SIMO channel estimation that follows the approximate maximum likelihood criterion, and propose an approximate ML estimation method.

The paper is organized as follows. In Section II we give the signal model of a time-slotted system and the low rank property of the channel matrix, and then derive the channel estimator by the spatial signal subspace projection in the context of maximum likelihood estimation. In Section III, we propose a new channel estimation method by the space-time signal subspace projection in the context of the ML channel estimation. In Section IV, we propose approximate ML channel estimation methods using soft information. The Cramer-Rao bounds are derived in Section V. The simulation results are given in Section VI. Section VII concludes the paper.

II. Problem Formulation

A. Signal Model

We here consider a time-slotted wireless communication system composed of a mobile user and a base station, where the base station employs an M-element antenna array. For the uplink case, a data frame consisting of the training sequence and the user's data sequence, which is unknown to the receiver, within each time slot is transmitted to the base station by the mobile user. The discrete baseband signal received at the based station array [6] is denoted by

$$\boldsymbol{y}(t) = \sum_i \boldsymbol{h}(i)x(t-i) + \boldsymbol{v}(t), \tag{1}$$

where $\boldsymbol{y}(t)$ denotes the $M \times 1$ received signal vector, $x(t)$ denotes the symbol transmitted at discrete time t [the discrete time interval is equal to the sampling period, or the period of the transmitted symbols], $\boldsymbol{h}(t)$ denotes the $M \times 1$ channel impulse response vector, $\boldsymbol{v}(t)$ is the $M \times 1$ array noise vector and Gaussian distributed with zero mean and the correlation matrix $\mathbb{E}[\boldsymbol{v}(t)\boldsymbol{v}^H(t-t_0)] = \boldsymbol{Q}\delta(t_0)$, and $\boldsymbol{v}(t)$ is assumed independent of $x(t)$. Notice that when $\boldsymbol{Q}$ is equal to $\sigma_v^2\boldsymbol{I}_M$[$\boldsymbol{I}_M$ denotes the $M \times M$ identity matrix], the array noise is considered spatially white, otherwise the array noise is colored with unknown interferences included. By assuming the channel length of W symbol periods, and letting $\boldsymbol{x}(t) \triangleq [x(t), x(t-1), \cdots, x(t-W+1)]^T$ and $\boldsymbol{H} \triangleq [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_W]$ with $\boldsymbol{h}_{i+1} \triangleq \boldsymbol{h}(i)$, the received signal (1) can be denoted by

$$\boldsymbol{y}(t) = \boldsymbol{H}\boldsymbol{x}(t) + \boldsymbol{v}(t), \tag{2}$$

where $\boldsymbol{H}$ denotes the $M \times W$ SIMO channel matrix and is assumed to be invariant within each slot. By denoting the training sequence as $\{x_t(i), i = 1, \ldots, \tilde{N}_t\}$ and the user data sequence as $\{x_d(i), i = 1, \ldots, \tilde{N}_d\}$ and assuming that a guard interval of $W-1$ symbol periods follows each sequence, all the samples $\boldsymbol{y}(t)$ within a slot can be represented by

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{V}, \tag{3}$$

or

$$\begin{cases} \boldsymbol{Y}_t = \boldsymbol{H}\boldsymbol{X}_t + \boldsymbol{V}_t, \\ \boldsymbol{Y}_d = \boldsymbol{H}\boldsymbol{X}_d + \boldsymbol{V}_d, \end{cases} \tag{4}$$

where $\boldsymbol{Y} = [\boldsymbol{Y}_t, \boldsymbol{Y}_d] = [\boldsymbol{y}(1), \cdots, \boldsymbol{y}(N_s)]$, $\boldsymbol{X} = [\boldsymbol{X}_t, \boldsymbol{X}_d]$, $\boldsymbol{V} = [\boldsymbol{V}_t, \boldsymbol{V}_d] = [\boldsymbol{v}(1), \cdots, \boldsymbol{v}(N_s)]$, $\boldsymbol{X}_t$ denotes the $W \times N_t$ training data matrix with the (m,n)th entry being $x_t(n-m+1)$, and $\boldsymbol{X}_d$ denotes the $W \times N_d$ data matrix with the (m,n)th entry being $x_d(n-m+1)$. Notice that $\boldsymbol{X}_t$ is Toeplitz. It is obvious that $N_t = \tilde{N}_t + W - 1$, $N_d = \tilde{N}_d + W - 1$, and $N_s = N_t + N_d = \tilde{N}_t + \tilde{N}_d + 2W - 2$.

B. Low Rank Channel Model

As shown in [6], the multipath channel can be modeled as a superposition of P resolvable path rays with the directions of arrival $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_P]$, the complex path gains $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_P]$, and the delays of path rays $\boldsymbol{\tau} = [\tau_1, \cdots, \tau_P]$. This means that the channel matrix $\boldsymbol{H}$ in (2) can be expressed as

$$\boldsymbol{H} = \sum_{p=1}^{P} \alpha_p \boldsymbol{a}(\theta_p)\boldsymbol{g}(\tau_p) = \boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{\Lambda}\boldsymbol{G}^T(\boldsymbol{\tau}), \tag{5}$$

where $\boldsymbol{F}(\boldsymbol{\theta}) = [\boldsymbol{a}(\theta_1), \cdots, \boldsymbol{a}(\theta_P)]$ denotes the array manifold, $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\alpha})$ denotes the diagonal matrix composed of the fading path gains, $\boldsymbol{G}(\boldsymbol{\tau}) = [\boldsymbol{g}(\tau_1), \cdots, \boldsymbol{g}(\tau_P)]$ denotes the delay pattern matrix with $\boldsymbol{g}(\tau_p) = [g(-\tau_p), g(T_s - \tau_p), \cdots, g((W-1)T_s - \tau_p)]^T$, and $g(t)$ represents pulse shaping function. Denoting $r_{max} = \min\{M, W\}$ and the rank of $\boldsymbol{H}$, $\boldsymbol{F}$ and $\boldsymbol{G}$ as r_0, r_s and r_t respectively, we have $r_0 = \min\{r_s, r_t\}$ according to (5). It is seen that when $r_0 < r_{max}$, $\boldsymbol{H}$ is rank-deficient and can be rewritten as the product of two full rank matrix [1] [5]:

$$\boldsymbol{H} = \boldsymbol{A}\boldsymbol{B}^H, \tag{6}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_{r_0}]$ is a $M \times r_0$ matrix and $\boldsymbol{B} = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_{r_0}]$ is a $W \times r_0$ matrix.

C. Channel Estimation by Signal Subspace Projection Using Training and User Data Sequences

The channel estimation by the signal subspace projection using training and user data sequences was proposed in [4], where the array noise is assumed spatially white and Gaussian distributed. Here we derive this channel estimation method in the sense of maximum likelihood. According to (3), the log-likelihood function of the received signal within a slot is given by

$$L(\boldsymbol{H},\boldsymbol{Q},\boldsymbol{X}_d) = -\log|\boldsymbol{Q}| - \frac{1}{N_s}\|\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}\|^2_{\boldsymbol{Q}^{-1}}, \quad (7)$$

where $\|\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}\|^2_{\boldsymbol{Q}^{-1}} = \mathrm{tr}[(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X})^H\boldsymbol{Q}^{-1}(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X})]$ and "tr" stands for trace of a matrix. As shown by [16], the optimal estimate of $\boldsymbol{Q}$ can be obtained by maximizing (7) for any given $\boldsymbol{H}$ and $\boldsymbol{X}$, and is given by the following solution when $\frac{1}{N_s}(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X})(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X})^H$ is positive definite:

$$\hat{\boldsymbol{Q}} = \frac{1}{N_s}(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X})(\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X})^H. \quad (8)$$

By substituting (8) and (6) in (7) and ignoring the constant, the log-likelihood function can be reduced to [17]:

$$\begin{aligned}
&L(\boldsymbol{A},\boldsymbol{B},\boldsymbol{X}_d)\\
&= -\log\left|(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{B}^H\boldsymbol{X})(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{B}^H\boldsymbol{X})^H\right|\\
&= -\log\Big|(\boldsymbol{Y}_t-\boldsymbol{A}\boldsymbol{B}^H\boldsymbol{X}_t)(\boldsymbol{Y}_t-\boldsymbol{A}\boldsymbol{B}^H\boldsymbol{X}_t)^H\\
&\quad +(\boldsymbol{Y}_d-\boldsymbol{A}\boldsymbol{B}^H\boldsymbol{X}_d)(\boldsymbol{Y}_d-\boldsymbol{A}\boldsymbol{B}^H\boldsymbol{X}_d)^H\Big|\\
&= -\log\Big|N_t\left(\boldsymbol{A}\tilde{\boldsymbol{B}}^H-\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}}\right)\left(\boldsymbol{A}\tilde{\boldsymbol{B}}^H-\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}}\right)^H\\
&\quad +N_t\boldsymbol{Q}_t+(\boldsymbol{Y}_d-\boldsymbol{A}\boldsymbol{B}^H\boldsymbol{X}_d)(\boldsymbol{Y}_d-\boldsymbol{A}\boldsymbol{B}^H\boldsymbol{X}_d)^H\Big|, \quad (9)
\end{aligned}$$

where $\tilde{\boldsymbol{B}} = \boldsymbol{R}_{x_t}^{\frac{1}{2}}\boldsymbol{B}$, $\boldsymbol{Q}_t = \boldsymbol{R}_{y_t} - \boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-1}\boldsymbol{R}_{c_t}^H$, $\boldsymbol{R}_{x_t} = \frac{1}{N_t}\boldsymbol{X}_t\boldsymbol{X}_t^H$, $\boldsymbol{R}_{y_t} = \frac{1}{N_t}\boldsymbol{Y}_t\boldsymbol{Y}_t^H$, $\boldsymbol{R}_{c_t} = \frac{1}{N_t}\boldsymbol{Y}_t\boldsymbol{X}_t^H$.

As shown by [17], $\boldsymbol{Q}_t \to \boldsymbol{Q}$ with probability 1 as $N_t \to \infty$, which means that $\boldsymbol{Q}_t \succ 0$ with probability 1 as $N_t \to \infty$. By defining $\boldsymbol{K} = \boldsymbol{X}_d^H\boldsymbol{B}$ and assuming $\boldsymbol{Q}_t \succ 0$, the log-likelihood function can be equivalently written as

$$\begin{aligned}
&L(\boldsymbol{A},\boldsymbol{B},\boldsymbol{K})\\
&= -\log|N_t\boldsymbol{Q}_t| - \log\left|\left(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{B}}^H-\tilde{\boldsymbol{H}}\right)\left(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{B}}^H-\tilde{\boldsymbol{H}}\right)^H\right.\\
&\quad \left.+\boldsymbol{I}+\frac{1}{N_t}(\tilde{\boldsymbol{Y}}_d-\tilde{\boldsymbol{A}}\boldsymbol{K}^H)(\tilde{\boldsymbol{Y}}_d-\tilde{\boldsymbol{A}}\boldsymbol{K}^H)^H\right|\\
&= -\log\left|\boldsymbol{I}+(\boldsymbol{Y}_A-\tilde{\boldsymbol{A}}\boldsymbol{B}_A^H)(\boldsymbol{Y}_A-\tilde{\boldsymbol{A}}\boldsymbol{B}_A^H)^H\right| - \log|N_t\boldsymbol{Q}_t|\\
&= -\log|N_t\boldsymbol{Q}_t| - \log|\boldsymbol{I}+\boldsymbol{\Phi}|, \quad (10)
\end{aligned}$$

where

$$\boldsymbol{Y}_A = [\tilde{\boldsymbol{H}}, \frac{1}{\sqrt{N_t}}\tilde{\boldsymbol{Y}}_d],\ \boldsymbol{B}_A^H = [\tilde{\boldsymbol{B}}^H, \frac{1}{\sqrt{N_t}}\boldsymbol{K}^H], \quad (11)$$

$$\begin{aligned}
\boldsymbol{\Phi} &= \left(\boldsymbol{Y}_A-\tilde{\boldsymbol{A}}\boldsymbol{B}_A^H\right)^H\left(\boldsymbol{Y}_A-\tilde{\boldsymbol{A}}\boldsymbol{B}_A^H\right)\\
&= \left[\boldsymbol{B}_A^H-\tilde{\boldsymbol{A}}^\dagger\boldsymbol{Y}_A\right]^H\tilde{\boldsymbol{A}}^H\tilde{\boldsymbol{A}}\left[\boldsymbol{B}_A^H-\tilde{\boldsymbol{A}}^\dagger\boldsymbol{Y}_A\right]+\boldsymbol{\Phi}_0, \quad (12)
\end{aligned}$$

$\tilde{\boldsymbol{H}} = \boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}}$, $\tilde{\boldsymbol{Y}}_d = \boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{Y}_d$, $\tilde{\boldsymbol{A}} = \boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{A}$, $\tilde{\boldsymbol{A}}^\dagger = (\tilde{\boldsymbol{A}}^H\tilde{\boldsymbol{A}})^{-1}\tilde{\boldsymbol{A}}^H$ being the pseudoinverse of $\tilde{\boldsymbol{A}}$, $\boldsymbol{\Phi}_0 = \boldsymbol{Y}_A^H\boldsymbol{Y}_A - \boldsymbol{Y}_A^H\boldsymbol{P}_{\tilde{\boldsymbol{A}}}\boldsymbol{Y}_A$, and $\boldsymbol{P}_{\tilde{\boldsymbol{A}}} = \tilde{\boldsymbol{A}}(\tilde{\boldsymbol{A}}^H\tilde{\boldsymbol{A}})^{-1}\tilde{\boldsymbol{A}}^H$ denoting the projection matrix onto the column space of $\tilde{\boldsymbol{A}}$. It is obvious that

$$\boldsymbol{\Phi} \succeq \boldsymbol{\Phi}_0 \quad (13)$$

and the equality holds if and only if

$$\boldsymbol{B}_A^H = \tilde{\boldsymbol{A}}^\dagger\boldsymbol{Y}_A. \quad (14)$$

This means that the log-likelihood function denoted by (10) can be maximized under the condition of (14). With the definition of (11), (14) can be separately denoted by

$$\tilde{\boldsymbol{B}}^H = \tilde{\boldsymbol{A}}^\dagger\tilde{\boldsymbol{H}} \Rightarrow \tilde{\boldsymbol{B}}^H = \tilde{\boldsymbol{A}}^\dagger\boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}}, \quad (15)$$

$$\boldsymbol{K}^H = \tilde{\boldsymbol{A}}^\dagger\tilde{\boldsymbol{Y}}_d, \quad (16)$$

where $\tilde{\boldsymbol{A}}$ needs to be determined. Notice that the rank of $\boldsymbol{P}_{\tilde{\boldsymbol{A}}}$ is r_0 and the rank of $\boldsymbol{\Phi}_0$ is $r_{max}-r_0$. By substituting (14) in (10), the maximum likelihood estimation (MLE) of $\tilde{\boldsymbol{A}}$ can be obtained by minimizing the following objective function

$$\log|\boldsymbol{I}+\boldsymbol{\Phi}_0| = \log\prod_{i=1}^{r_{max}-r_0}(1+\lambda_i(\boldsymbol{\Phi}_0)), \quad (17)$$

where $\{\lambda_i(\boldsymbol{\Phi}_0)\}_{i=1}^{r_{max}-r_0}$ denote the eigenvalues of $\boldsymbol{\Phi}_0$ in descending order. Following the Poincaré separation theorem [18], we can have

$$\lambda_i(\boldsymbol{Y}_A\boldsymbol{Y}_A^H) \geq \lambda_i(\boldsymbol{\Phi}_0) \geq \lambda_{r_0+i}(\boldsymbol{Y}_A\boldsymbol{Y}_A^H), \quad (18)$$

where the second equality holds if and only if $\boldsymbol{P}_{\tilde{\boldsymbol{A}}} = \boldsymbol{U}_{\tilde{\boldsymbol{A}}}\boldsymbol{U}_{\tilde{\boldsymbol{A}}}^H$, with $\boldsymbol{U}_{\tilde{\boldsymbol{A}}}$ denoting the matrix by the eigenvectors corresponding to r_0 leading eigenvalues of $\boldsymbol{Y}_A\boldsymbol{Y}_A^H$, *i.e.*, $\{\lambda_i(\boldsymbol{Y}_A\boldsymbol{Y}_A^H)\}_{i=1}^{r_0}$. r_0 can be estimated using Akaike information criterion (AIC) [20] or minimum description length (MDL) [19] criterion with the eigenvalues of $\tilde{\boldsymbol{H}}\tilde{\boldsymbol{H}}^H$. With $\boldsymbol{P}_{\tilde{\boldsymbol{A}}}$ and substituting (15) in (6), the channel estimate can be obtained by the signal subspace projection:

$$\hat{\boldsymbol{H}}_s = \boldsymbol{A}\boldsymbol{B}^H = \boldsymbol{Q}_t^{\frac{H}{2}}\boldsymbol{P}_{\tilde{\boldsymbol{A}}}\boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-1}. \quad (19)$$

According to the definition of $\boldsymbol{Q}_t$ given in (9) and substituting $\boldsymbol{Y}_t$ in $\boldsymbol{Q}_t$, we can derive

$$\boldsymbol{Q}_t = \frac{1}{N_t}\left[\boldsymbol{V}_t\boldsymbol{V}_t^H - \boldsymbol{V}_t\boldsymbol{X}_t^H(\boldsymbol{X}_t\boldsymbol{X}_t^H)^{-1}\boldsymbol{X}_t\boldsymbol{V}_t^H\right], \quad (20)$$

$$\begin{aligned}
\mathbb{E}\big(\boldsymbol{Q}_t(i,j)\big) &= \frac{1}{N_t}\mathbb{E}\left\{\boldsymbol{v}_{ti}\boldsymbol{v}_{tj}^H - \mathrm{tr}\left[\boldsymbol{X}_t^H(\boldsymbol{X}_t\boldsymbol{X}_t^H)^{-1}\boldsymbol{X}_t\boldsymbol{v}_{tj}^H\boldsymbol{v}_{ti}\right]\right\}\\
&= \boldsymbol{Q}(i,j)\left\{1-\frac{1}{N_t}\mathrm{tr}\left[\boldsymbol{X}_t^H(\boldsymbol{X}_t\boldsymbol{X}_t^H)^{-1}\boldsymbol{X}_t\right]\right\}\\
&= \boldsymbol{Q}(i,j)(1-\frac{W}{N_t}), \quad (21)
\end{aligned}$$

where $\mathbb{E}(\cdot)$ denotes expectation, $\boldsymbol{Q}_t(i,j)$ the (i,j)-th element of $\boldsymbol{Q}_t$, and $\boldsymbol{v}_{ti}$ the ith row of $\boldsymbol{V}_t$. With the result shown above, it is obvious that the estimate of $\boldsymbol{Q}$ can be denoted by

$$\hat{\boldsymbol{Q}} \doteq \frac{N_t}{N_t-W}\boldsymbol{Q}_t. \quad (22)$$

When all the transmitted symbols are assumed to be known, *i.e.*, $N_d = 0$, (19) is reduced to that of the RR [5], which does not employ the data sequence in channel estimation. Moreover, we note that (19) is equivalent to projecting the prewhitened LS estimate of the channel matrix to its spatial signal subspace obtained by employing both training and the data sequences within each slot. When $M \leq W$ and the number of resolvable path rays is larger than that of the array elements, *i.e.*, $r_0 = M$, the channel estimate by (19) is reduced to

$$\hat{\boldsymbol{H}}_s = \hat{\boldsymbol{H}} = \boldsymbol{R}_{c_t} \boldsymbol{R}_{x_t}^{-1}, \tag{23}$$

which coincides with the LS estimate. In this case, the spatial signal subspace projection performed in (19) does not make sense.

III. Channel Estimation Based on Space-time Signal Subspace Projection

A. Channel Estimation by Space-Time Signal Subspace Projection

Since the number of array elements is usually smaller than the maximum length of the channel in symbol periods, it is appropriate in practice to assume that $M \leq W$. For example, the scenario mentioned in [22] assumes that $M \leq 8$ while W could be as large as $W_{max} = 57T_s$ [T_s denotes symbol period]. By stacking N contiguous discrete signal vectors of (2), the received signal represented by a space-time signal vector is denoted by

$$\begin{aligned} \boldsymbol{y}_E(t) &= [\boldsymbol{y}^T(t), \boldsymbol{y}^T(t-1), \cdots, \boldsymbol{y}^T(t-N+1)]^T \\ &= \mathcal{H}\boldsymbol{x}_E(t) + \boldsymbol{v}_E(t), \end{aligned} \tag{24}$$

where

$$\begin{aligned} \boldsymbol{x}_E(t) &= [x(t), x(t-1), \cdots, x(t-N-W+2)]^T, \\ \boldsymbol{v}_E(t) &= [\boldsymbol{v}^T(t), \boldsymbol{v}^T(t-1), \cdots, \boldsymbol{v}^T(t-N+1)]^T, \end{aligned} \tag{25}$$

and $\mathcal{H}$ denotes a $MN \times (W+N-1)$ Toeplitz channel matrix by the matrix function $S_N(\cdot)$ for simplicity of notation:

$$\begin{aligned} \mathcal{H} &= S_N(\boldsymbol{H}) \\ &= \begin{bmatrix} \boldsymbol{h}_1 & \cdots & \boldsymbol{h}_W & 0 & \cdots & 0 \\ 0 & \boldsymbol{h}_1 & \cdots & \boldsymbol{h}_W & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{h}_1 & \cdots & \boldsymbol{h}_W \end{bmatrix}. \end{aligned} \tag{26}$$

This function transforms a matrix to its corresponding block Toeplitz matrix.

By using multiple samples of the received signal vector, we can obtain the space-time correlation matrix. Eigendecomposing this correlation matrix will yield the space-time signal subspace, which is spanned by the principal eigenvectors of the matrix. Since this signal subspace is intrinsically the column space of $\mathcal{H}$, the condition for the existence of this signal subspace is that $\mathcal{H}$ must be full column rank, or the rank of $\mathcal{H}$ must be less than MN. To meet such a condition, it is necessary to choose a suitable value for N. In the following theorem we prove a lower bound of N to meet the condition.

Theorem 1: When $M < W$ and the number of resolvable path rays is larger than that of the array elements, *i.e.*, $r_0 = r_s = M$, the minimum value of N selected for the column rank of $\mathcal{H}$ to be less than MN is $N_0 = \lfloor \frac{W-1}{M-1} \rfloor + 1$.

Proof: According to (5), $\mathcal{H}$ can be denoted as $\mathcal{H} = \boldsymbol{F}_N \boldsymbol{\Lambda}_N \mathcal{G}$, where $\boldsymbol{F}_N = \boldsymbol{I}_N \otimes \boldsymbol{F}(\boldsymbol{\theta})$, $\boldsymbol{\Lambda}_N = \boldsymbol{I}_N \otimes \boldsymbol{\Lambda}$, $\mathcal{G} = S_N(\boldsymbol{G}^T(\boldsymbol{\tau}))$ and "$\otimes$" denotes Kronecker product. When $r_s = M$, the rank of these three matrices are that $r_S = \text{rank}(\boldsymbol{F}_N) = MN$, $\text{rank}(\boldsymbol{\Lambda}_N) = PN$, and $r_T = \text{rank}(\mathcal{G}) \leq \min\{PN, W+N-1\}$. Since $P \geq M$, the rank of $\mathcal{H}$, denoted as $r_{\mathcal{H}}$, is equal to $\min\{r_S, r_T\}$. Denoting $N_0 = \lfloor \frac{W-1}{M-1} \rfloor + 1$ and choosing $N \geq N_0$, we can have $W + N - 1 < MN$ and straightforwardly $r_{\mathcal{H}} < MN$. ■

As W is always chosen to be larger than the real length of channel response to avoid interference from adjacent slots, a smaller N, *e.g.*, $N = \lfloor \frac{W'-1}{M-1} \rfloor + 1$ with $W' < W$, can be used to reduce the dimension of the signal subspace and the complexity of the method. Notice that larger N will lead to greater dimension of noise subspace, which may not be helpful to significantly improve the channel estimation performance.

With properly chosen N, the Toeplitz channel matrix can be represented by the product of two full column rank matrices in the same manner as that of (6)

$$\mathcal{H} = \mathcal{A}\mathcal{B}^H, \tag{27}$$

where $\mathcal{A}$ is the $MN \times r_{\mathcal{H}}$ matrix and $\mathcal{B}$ is with the dimension of $(W+N-1) \times r_{\mathcal{H}}$.

According to (24), the signal received within a slot can be represented by

$$\begin{cases} \mathcal{Y}_t = S_N(\boldsymbol{H}\boldsymbol{X}_t) + \mathcal{V}_t = \mathcal{H}\boldsymbol{X}_T + \mathcal{V}_t, \\ \mathcal{Y}_d = S_N(\boldsymbol{H}\boldsymbol{X}_d) + \mathcal{V}_d = \mathcal{H}\boldsymbol{X}_D + \mathcal{V}_d, \end{cases} \tag{28}$$

where

$$\begin{aligned} \mathcal{Y}_t &= \begin{bmatrix} \boldsymbol{y}(1) & \cdots & \boldsymbol{y}(N_t) & 0 & \cdots & 0 \\ 0 & \boldsymbol{y}(1) & \cdots & \boldsymbol{y}(N_t) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{y}(1) & \cdots & \boldsymbol{y}(N_t) \end{bmatrix} \\ &= S_N(\boldsymbol{Y}_t) \end{aligned} \tag{29}$$

denotes the $MN \times \check{N}_t$ block Toeplitz matrix of the received signal corresponding to the training sequence, $\boldsymbol{X}_T$ is the Toeplitz training data matrix with its dimension of $(W+N-1) \times \check{N}_t$ and its (m,n)th entry being $x_t(n-m+1)$, $\mathcal{V}_t = S_N(\boldsymbol{V}_t)$ denotes the block Toeplitz noise matrix, and similarly $\mathcal{Y}_d = S_N(\boldsymbol{Y}_d)$ with the dimension $MN \times \check{N}_d$, the data matrix $\boldsymbol{X}_D$ with the dimension $(W+N-1) \times \check{N}_t$, and $\mathcal{V}_d = S_N(\boldsymbol{V}_d)$. Notice that $\check{N}_t = N_t + N - 1$, and $\check{N}_d = N_d + N - 1$.

By denoting $\mathcal{Y}_A = [\tilde{\mathcal{H}}, \tilde{\mathcal{Y}}_d]$, $\tilde{\mathcal{H}} = S_N(\frac{1}{\sqrt{N_t}} \boldsymbol{Q}_t^{-\frac{H}{2}} \hat{\boldsymbol{H}}) \boldsymbol{X}_T$, $\tilde{\mathcal{Y}}_d = \frac{1}{\sqrt{N_t}} (\boldsymbol{I}_N \otimes \boldsymbol{Q}_t^{-\frac{H}{2}}) \mathcal{Y}_d$, and $\boldsymbol{P}_{\tilde{\mathcal{A}}}$ as the projection matrix of the subspace spanned by the $r_{\mathcal{H}}$ principle eigenvectors of $\mathcal{Y}_A \mathcal{Y}_A^H$, the channel estimation by space-time signal subspace projection can be performed by

$$\hat{\mathcal{H}}_S = (\boldsymbol{I}_N \otimes \boldsymbol{Q}_t^{\frac{H}{2}}) \boldsymbol{P}_{\tilde{\mathcal{A}}} S_N(\boldsymbol{Q}_t^{-\frac{H}{2}} \hat{\boldsymbol{H}}). \tag{30}$$

The derivation of (30) is given in Appendix A.

Denoting the (i,j)-th block entry of $\tilde{\mathcal{H}}_S$ by the $M\times 1$ vector $\tilde{\boldsymbol{h}}_{i,j}$ and employing the block Toeplitz structure of $\mathcal{H}$, the new estimate of $\boldsymbol{H}$ can be written as

$$\hat{\boldsymbol{H}}_{st} = \frac{1}{N}\sum_{t=1}^{N}([\tilde{\boldsymbol{h}}_{t,t}, \tilde{\boldsymbol{h}}_{t,t+1}, \cdots, \tilde{\boldsymbol{h}}_{t,t+W-1}]). \quad (31)$$

Notice that in comparison with the method of LS channel estimation, the additional computational burden by the space-time signal subspace projection lies at the computation of the correlation matrix $\mathcal{Y}_A\mathcal{Y}_A^H$, which is with the order of $\mathcal{O}(M^2N^2N_d)$, and its corresponding eigendecomposition for eigenvalues and eigenvectors has the complexity of $\mathcal{O}(M^3N^3)$.

Furthermore, it is worthy to mention that some subspace methods based on stacking multiple contiguous observations and combining training and data signals have been already proposed in [14] and [15], where the estimation is performed on vectorized form of the channel matrix, i.e., $\boldsymbol{h} \triangleq \mathrm{vec}(\boldsymbol{H}) = [\boldsymbol{h}_1^T, \cdots, \boldsymbol{h}_W^T]^T$. However, in these methods, the problem of finding the optimal trade-off between the training sequence based criterion and the blind subspace (of user data) criterion is nontrivial and must be considered. Instead, this problem can be circumvented by directly estimating the block Toeplitz channel matrix.

B. Channel Estimation Using Multiple Slots

In the cases where the variation of the angle/delay pattern $\{\boldsymbol{\theta}, \boldsymbol{\tau}\}$ over L slots are smaller than the angular-temporal resolution [6], and the fading path gains $\boldsymbol{\alpha}$ that change randomly slot by slot have the correlation matrix $\boldsymbol{R}_{\boldsymbol{\alpha}} = \mathbb{E}(\boldsymbol{\alpha}\boldsymbol{\alpha}^H) = \mathrm{diag}\{[\sigma_1^2, \cdots, \sigma_P^2]\}$, the channel matrix of the l-th slot can be parameterized by [6]

$$\boldsymbol{H}(l) = \boldsymbol{U}_s\Gamma(l)\boldsymbol{U}_t^H, \quad (32)$$

where $\boldsymbol{U}_s$ is the $M \times r_s(L)$ matrix, $\boldsymbol{U}_t^H$ is the $W \times r_t(L)$ matrix, both are full rank and remain constant within L slots, and $\Gamma(l)$ is the $r_s(L) \times r_t(L)$ full-rank matrix that varies between slots.

With the use of all received samples, the multi-slot channel estimation can be obtained by

$$\hat{\boldsymbol{H}}_{mst}(l) = \boldsymbol{Q}_t^{\frac{H}{2}}\boldsymbol{P}_{\tilde{U}_s}\tilde{\boldsymbol{H}}(l)\boldsymbol{P}_{\tilde{U}_t}\boldsymbol{R}_{x_t}^{-\frac{H}{2}}, \quad (33)$$

where $\tilde{\boldsymbol{H}}(l) = \boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{R}_{c_t}(l)\boldsymbol{R}_{x_t}^{-\frac{1}{2}}$, $\boldsymbol{P}_{\tilde{U}_s}$ and $\boldsymbol{P}_{\tilde{U}_t}$ are projection matrices of the spatial and temporal subspace defined by (77) and (86) in Appendix B, respectively.

The derivation of (33) is given in Appendix B.

When $r_s = M$, it is evident to see that $\boldsymbol{P}_{\tilde{U}_s} = \boldsymbol{I}$, which means that this projection does not make sense. In this case, combining the channel estimation by space-time signal subspace projection proposed in the previous subsection with the idea of the abovementioned multi-slot channel estimation will lead to a new channel estimate:

$$\hat{\mathcal{H}}_{mstt}(l) = (\boldsymbol{I}_N \otimes \boldsymbol{Q}_t^{\frac{H}{2}})\boldsymbol{P}_{\hat{\mathcal{A}}}(l)S_N(\tilde{\boldsymbol{H}}(l)\boldsymbol{P}_{\tilde{U}_t}\boldsymbol{R}_{x_t}^{-\frac{H}{2}}), \quad (34)$$

where $\boldsymbol{P}_{\hat{\mathcal{A}}}(l)$ is the projection matrix of the subspace spanned by the $r_{\mathcal{H}}(l)$ principle eigenvectors of $\mathcal{Y}_A(l)\mathcal{Y}_A^H(l)$. Notice that the channel estimate $\hat{\boldsymbol{H}}_{mstt}(l)$ can be obtained from $\hat{\mathcal{H}}_{mstt}(l)$ in the same manner as (31).

IV. Channel Estimation Using Soft Information

A. Approximated ML SIMO Channel Estimation

We here consider the problem of SIMO channel estimation using the soft information. As shown by Fig. 1, in the framework of a turbo receiver, the soft information is obtained from the maximum *a posteriori* probability (MAP) decoder and fed back for soft cancellation and minimum mean squared error (SC-MMSE) equalization and channel estimation. According to the signal model for soft iterative channel estimation [13], the log-likelihood ratios (LLRs) $\lambda_1[b(t)]$ on the interleaved coded bits $b(t)$ can be exploited to refine the channel estimates iteration by iteration. In the case that Gray mapping is used for QPSK modulation, we have [8]

$$\begin{aligned}
&\bar{x}_d(t) = \frac{1}{\sqrt{2}}\left(\tanh\frac{\lambda_1[b(2t)]}{2} + i\tanh\frac{\lambda_1[b(2t+1)]}{2}\right),\\
&\Delta x_d(t) = x_d(t) - \bar{x}_d(t),\ t = 1, \cdots, \tilde{N}_d,\\
&\sigma_d^2(t) = \mathbb{E}|\Delta x_d(t)|^2 = 1 - |\bar{x}_d(t)|^2,
\end{aligned}$$

where $\lambda_1[b(t)] = \ln\frac{P(b(t)=+1)}{P(b(t)=-1)}$.

Define $\bar{\boldsymbol{X}}_d = \mathbb{E}(\boldsymbol{X}_d)$, $\Delta\boldsymbol{X}_d = \boldsymbol{X}_d - \bar{\boldsymbol{X}}_d$, and $\Delta\boldsymbol{V}_d = \boldsymbol{H}\Delta\boldsymbol{X}_d = [\Delta\boldsymbol{v}_d(1), \cdots, \Delta\boldsymbol{v}_d(N_d)]$. The received signal matrix (3) can be represented by two parts:

$$\begin{cases}\boldsymbol{Y}_t = \boldsymbol{H}\boldsymbol{X}_t + \boldsymbol{V}_t,\\ \boldsymbol{Y}_d = \boldsymbol{H}\bar{\boldsymbol{X}}_d + \Delta\boldsymbol{V}_d + \boldsymbol{V}_d.\end{cases} \quad (35)$$

Here we assume that $\boldsymbol{v}(t)$ is spatially white with $\boldsymbol{Q} = \sigma_v^2\boldsymbol{I}_M$, the user data symbols $\{x_d(t)\}$ are independent so that $\boldsymbol{X}_d\boldsymbol{X}_d^H \approx \tilde{N}_d\boldsymbol{I}_W$, and the sequence $\{\Delta x_d(t)\}$ is a stationary white noise process with the zero mean and the variance σ_d^2 ($\sigma_d^2 = \frac{1}{\tilde{N}_d}\sum_t\sigma_d^2(t)$). These assumptions lead to that $\mathbb{E}[\Delta\boldsymbol{x}_d(t)\Delta\boldsymbol{x}_d^H(t)] = \sigma_d^2\boldsymbol{I}$, $\mathbb{E}[\Delta\boldsymbol{v}_d(t)] = 0$, and $\mathbb{E}[\Delta\boldsymbol{v}_d(t)\Delta\boldsymbol{v}_d^H(t)] = \sigma_d^2\boldsymbol{R}_H$, where $\boldsymbol{R}_H = \boldsymbol{H}\boldsymbol{H}^H$. By assuming that $\Delta\boldsymbol{v}_d(t)$ follows the $M\times 1$ complex Gaussian distribution, the approximate ML channel estimation can be performed by maximizing the log-likelihood function

$$\begin{aligned}\hat{\boldsymbol{H}} = \arg\max_{\boldsymbol{H}}\ \Big[&-N_t\log|\sigma_v^2\boldsymbol{I}| - N_d\log|\sigma_v^2\boldsymbol{I} + \sigma_d^2\boldsymbol{R}_H| - \frac{1}{\sigma_v^2}\\ &\times\|\boldsymbol{Y}_t - \boldsymbol{H}\boldsymbol{X}_t\|_F^2 - \|\boldsymbol{Y}_d - \boldsymbol{H}\bar{\boldsymbol{X}}_d\|^2_{(\sigma_v^2\boldsymbol{I}+\sigma_d^2\boldsymbol{R}_H)^{-1}}\Big].\end{aligned} \quad (36)$$

It is seen that the objective function of (36) contains a log-det function, which makes the problem nonlinear and non-convex. In the following, we aim to solve this problem by approximating $\boldsymbol{R}_H$ in the two ways.

In the first way, we assume that $\boldsymbol{R}_H = \rho\sigma_v^2\boldsymbol{I}$ with $\rho = \sigma_h^2/\sigma_v^2$ denoting the antenna element SNR, the problem (36) becomes

$$\hat{\boldsymbol{H}} = \arg\max_{\boldsymbol{H}}\left(-\frac{\|\boldsymbol{Y}_t - \boldsymbol{H}\boldsymbol{X}_t\|_F^2}{\sigma_v^2} - \frac{\|\boldsymbol{Y}_d - \boldsymbol{H}\bar{\boldsymbol{X}}_d\|_F^2}{\sigma_v^2 + \rho\sigma_v^2\sigma_d^2}\right). \quad (37)$$

It is obvious that the problem (37) can be analytically solved with the closed form solution denoted by

$$\hat{\boldsymbol{H}}_{s1} = \frac{1}{N_s}(N_t\boldsymbol{R}_{c_t} + N_d\gamma\bar{\boldsymbol{R}}_{c_d})\boldsymbol{R}_{td}^{-1}, \tag{38}$$

where $\gamma = (1+\sigma_d^2\rho)^{-1}$, $\bar{\boldsymbol{R}}_{c_d} = \frac{1}{N_d}\boldsymbol{Y}_d\bar{\boldsymbol{X}}_d^H$, $\boldsymbol{R}_{td} = (N_t\boldsymbol{R}_{x_t} + N_d\gamma\bar{\boldsymbol{R}}_{x_d})/N_s$, $\bar{\boldsymbol{R}}_{x_d} = \frac{1}{N_d}\bar{\boldsymbol{X}}_d\bar{\boldsymbol{X}}_d^H$, and the "s" in the subscript of $\hat{\boldsymbol{H}}_{s1}$ stands for soft-based estimation. It is seen that this method is similar to the method by using soft information in [13] except the additional weighting factor γ. Notice that γ is often less than or equal to 1, and the equality holds with perfect *a priori* information.

In the second way, we replace $\boldsymbol{R}_H$ by its estimate $\hat{\boldsymbol{R}}_H = \hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^H$, where $\hat{\boldsymbol{H}}$ is the channel estimate obtained in the previous iteration. The problem (36) thus becomes

$$\hat{\boldsymbol{H}} = \arg\min_{\boldsymbol{H}} \ \left(\|\boldsymbol{Y}_t - \boldsymbol{H}\boldsymbol{X}_t\|_F^2 + \|\boldsymbol{Y}_d - \boldsymbol{H}\bar{\boldsymbol{X}}_d\|_{\boldsymbol{G}}^2\right) \tag{39}$$

and $\hat{\boldsymbol{H}}$ is obtained by letting $dL(\boldsymbol{H})/d\boldsymbol{H} = 0$ [$L(\boldsymbol{H})$ is the objective function of (39)], which leads to

$$N_t\boldsymbol{H}\boldsymbol{R}_{x_t} + N_d\boldsymbol{G}\boldsymbol{H}\bar{\boldsymbol{R}}_{x_d} = \boldsymbol{C}, \tag{40}$$

where $\boldsymbol{G} = (\boldsymbol{I} + \frac{\sigma_d^2}{\sigma_v^2}\hat{\boldsymbol{R}}_H)^{-1}$ and $\boldsymbol{C} = N_t\boldsymbol{R}_{c_t} + N_d\boldsymbol{G}\bar{\boldsymbol{R}}_{c_d}^H = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_W]$.

By defining $\boldsymbol{c} = \left[\boldsymbol{c}_1^T, \boldsymbol{c}_2^T, \cdots, \boldsymbol{c}_W^T\right]^T$ and using *Lemma* 4.3.1 in [25], (40) can be equivalently rewritten as

$$(N_t\boldsymbol{R}_{x_t}^T \otimes \boldsymbol{I} + N_d\bar{\boldsymbol{R}}_{x_d}^T \otimes \boldsymbol{G})\boldsymbol{h} = \boldsymbol{c}, \tag{41}$$

where $\boldsymbol{h} \triangleq \mathrm{vec}(\boldsymbol{H})$.

Notice that $\boldsymbol{R}_{x_t}$, $\bar{\boldsymbol{R}}_{x_d}$, and $\boldsymbol{G}$ are always positive definite. This means that $N_t\boldsymbol{R}_{x_t}^T \otimes \boldsymbol{I} + N_d\bar{\boldsymbol{R}}_{x_d}^T \otimes \boldsymbol{G}$ is also positive definite and hence (41) has a unique solution. We denote this unique solution as

$$\hat{\boldsymbol{h}}_{s2} = (N_t\boldsymbol{R}_{x_t}^T \otimes \boldsymbol{I} + N_d\bar{\boldsymbol{R}}_{x_d}^T \otimes \boldsymbol{G})^{-1}\boldsymbol{c}. \tag{42}$$

According to the abovementioned definition of $\boldsymbol{h}$, the channel matrix estimate $\hat{\boldsymbol{H}}_{s2}$ can be straightforwardly formed by the terms of $\hat{\boldsymbol{h}}_{s2}$.

Notice that the knowledge of $\boldsymbol{R}_H$ is usually not available. This means that assuming known $\boldsymbol{R}_H$ may cause performance loss with this ML channel estimator.

B. Approximate ML Channel Estimation Based on Semidefinite Relaxation

To avoid the abovementioned problem, we here use the first-order Taylor series expansion to approximate the log-det function in (36), and then approach this ML channel estimation problem by semidefinite relaxations (SDR).

Define

$$\boldsymbol{Q}_s \triangleq \sigma_v^2\boldsymbol{I} + \sigma_d^2\boldsymbol{R}_H.$$

According to the derivation in [26], the first-order Taylor series expansion of $\log|\boldsymbol{Q}_s|$ *w.r.t* $\hat{\boldsymbol{R}}_H$ is given by

$$\begin{aligned}\log|\boldsymbol{Q}_s| \approx &\mathrm{tr}\left[(\sigma_v^2\boldsymbol{I} + \sigma_d^2\hat{\boldsymbol{R}}_H)^{-1}(\sigma_d^2\boldsymbol{R}_H - \sigma_d^2\hat{\boldsymbol{R}}_H)\right] \\ &+ \log|\sigma_v^2\boldsymbol{I} + \sigma_d^2\hat{\boldsymbol{R}}_H|.\end{aligned} \tag{43}$$

By omitting the constants and substituting (43) in (36), the ML SIMO channel estimation can be approximately performed by

$$\begin{aligned}\hat{\boldsymbol{H}} = \arg\max_{\boldsymbol{H}} \ &\Big\{-N_d\mathrm{tr}[(\sigma_v^2\boldsymbol{I} + \sigma_d^2\hat{\boldsymbol{R}}_H)^{-1}\sigma_d^2\boldsymbol{R}_H] - \frac{1}{\sigma_v^2} \\ &\times\|\boldsymbol{Y}_t - \boldsymbol{H}\boldsymbol{X}_t\|_F^2 - \|\boldsymbol{Y}_d - \boldsymbol{H}\bar{\boldsymbol{X}}_d\|_{\boldsymbol{Q}_s^{-1}}^2\Big\}.\end{aligned} \tag{44}$$

Notice that

$$\|\boldsymbol{Y}_t - \boldsymbol{H}\boldsymbol{X}_t\|_F^2 = N_t\|\boldsymbol{H}\boldsymbol{R}_{x_t}^{\frac{H}{2}} - \boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}}\|_F^2 + N_t\mathrm{tr}(\boldsymbol{Q}_t), \tag{45}$$

and

$$\begin{aligned}&\|\boldsymbol{Y}_d - \boldsymbol{H}\bar{\boldsymbol{X}}_d\|_{\boldsymbol{Q}_s^{-1}}^2 \\ =&N_d\mathrm{tr}\left[\boldsymbol{Q}_s^{-1}(\boldsymbol{H}\bar{\boldsymbol{R}}_{x_d}\boldsymbol{H}^H - \bar{\boldsymbol{R}}_{c_d}\boldsymbol{H}^H - \boldsymbol{H}\bar{\boldsymbol{R}}_{c_d}^H + \boldsymbol{R}_{y_d})\right] \\ =&N_d\|\boldsymbol{H}\bar{\boldsymbol{R}}_{x_d}^{\frac{H}{2}} - \bar{\boldsymbol{R}}_{c_d}\bar{\boldsymbol{R}}_{x_d}^{-\frac{1}{2}}\|_{\boldsymbol{Q}_s^{-1}}^2 + N_d\mathrm{tr}\left(\boldsymbol{Q}_s^{-1}\boldsymbol{R}_{vd}\right),\end{aligned} \tag{46}$$

where $\boldsymbol{R}_{y_d} = \frac{1}{N_d}\boldsymbol{Y}_d\boldsymbol{Y}_d^H$, $\boldsymbol{R}_{vd} = \boldsymbol{R}_{y_d} - \bar{\boldsymbol{R}}_{c_d}\bar{\boldsymbol{R}}_{x_d}^{-1}\bar{\boldsymbol{R}}_{c_d}^H$, $\boldsymbol{Q}_t$ is defined in (9), and the second term in (45) is a constant.

By substituting (45) and (46) in (44) and ignoring the constant term, (44) can be equivalently written as

$$\begin{aligned}\hat{\boldsymbol{H}} = \arg\max_{\boldsymbol{H}} \ \Big[&- N_d\mathrm{tr}(\boldsymbol{R}_0\boldsymbol{R}_H) - \frac{N_t}{\sigma_v^2}\|\boldsymbol{H}\boldsymbol{R}_{x_t}^{\frac{H}{2}} - \boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}}\|_F^2 \\ &- N_d\|\boldsymbol{H}\bar{\boldsymbol{R}}_{x_d}^{\frac{H}{2}} - \bar{\boldsymbol{R}}_{c_d}\bar{\boldsymbol{R}}_{x_d}^{-\frac{1}{2}}\|_{\boldsymbol{Q}_s^{-1}}^2 - N_d\mathrm{tr}\left(\boldsymbol{Q}_s^{-1}\boldsymbol{R}_{vd}\right)\Big],\end{aligned} \tag{47}$$

where $\boldsymbol{R}_0 = \sigma_d^2(\sigma_v^2\boldsymbol{I} + \sigma_d^2\hat{\boldsymbol{R}}_H)^{-1}$. Furthermore, (47) can be equivalently rewritten as the following nonlinear and nonconvex optimization problem:

$$\begin{aligned}&\min_{\boldsymbol{H},\boldsymbol{T}_1,\boldsymbol{T}_2,\boldsymbol{T}_3} \ \left\{N_t\mathrm{tr}(\boldsymbol{T}_1) + N_d[\mathrm{tr}(\boldsymbol{R}_0\boldsymbol{R}_H) + \mathrm{tr}(\boldsymbol{T}_2 + \boldsymbol{T}_3)]\right\} \\ &\text{subject to} \\ &(\boldsymbol{H}\boldsymbol{R}_{x_t}^{\frac{H}{2}} - \boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}})\frac{1}{\sigma_v^2}(\boldsymbol{H}\boldsymbol{R}_{x_t}^{\frac{H}{2}} - \boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}})^H \preceq \boldsymbol{T}_1, \\ &(\boldsymbol{H}\bar{\boldsymbol{R}}_{x_d}^{\frac{H}{2}} - \bar{\boldsymbol{R}}_{c_d}\bar{\boldsymbol{R}}_{x_d}^{-\frac{1}{2}})^H\boldsymbol{Q}_s^{-1}(\boldsymbol{H}\bar{\boldsymbol{R}}_{x_d}^{\frac{H}{2}} - \bar{\boldsymbol{R}}_{c_d}\bar{\boldsymbol{R}}_{x_d}^{-\frac{1}{2}}) \preceq \boldsymbol{T}_2, \\ &\boldsymbol{R}_{vd}^{\frac{1}{2}}\boldsymbol{Q}_s^{-1}\boldsymbol{R}_{vd}^{\frac{1}{2}} \preceq \boldsymbol{T}_3, \\ &\boldsymbol{R}_H = \boldsymbol{H}\boldsymbol{H}^H, \ \boldsymbol{Q}_s = \sigma_v^2\boldsymbol{I} + \sigma_d^2\boldsymbol{R}_H, \\ &\boldsymbol{T}_i \succeq 0, \ \boldsymbol{T}_i = \boldsymbol{T}_i^H, \ i = 1, 2, 3.\end{aligned} \tag{48}$$

Using the relaxation $\boldsymbol{R}_H = \boldsymbol{H}\boldsymbol{H}^H \Rightarrow \boldsymbol{H}\boldsymbol{H}^H \preceq \boldsymbol{R}_H$ and following Schur complement representation:

$$\boldsymbol{H}\boldsymbol{H}^H \preceq \boldsymbol{R}_H \Leftrightarrow \begin{bmatrix}\boldsymbol{I} & \boldsymbol{H}^H \\ \boldsymbol{H} & \boldsymbol{R}_H\end{bmatrix} \succeq 0, \tag{49}$$

the optimization problem (48) can be reformulated into the

following semidefinite programming (SDP):

$$\min_{\boldsymbol{H},\boldsymbol{T}_1,\boldsymbol{T}_2,\boldsymbol{T}_3,\boldsymbol{R}_H} \left\{N_t\text{tr}(\boldsymbol{T}_1)+N_d[\text{tr}(\boldsymbol{R}_0\boldsymbol{R}_H)+\text{tr}(\boldsymbol{T}_2+\boldsymbol{T}_3)]\right\}$$

subject to

$$\begin{bmatrix} \sigma_v^2\boldsymbol{I}_W & (\boldsymbol{H}\boldsymbol{R}_{x_t}^{\frac{H}{2}}-\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}})^H \\ \boldsymbol{H}\boldsymbol{R}_{x_t}^{\frac{H}{2}}-\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}} & \boldsymbol{T}_1 \end{bmatrix}\succeq 0,$$

$$\begin{bmatrix} \sigma_v^2\boldsymbol{I}+\sigma_d^2\boldsymbol{R}_H & \boldsymbol{H}\bar{\boldsymbol{R}}_{x_d}^{\frac{H}{2}}-\bar{\boldsymbol{R}}_{c_d}\bar{\boldsymbol{R}}_{x_d}^{-\frac{1}{2}} \\ (\boldsymbol{H}\bar{\boldsymbol{R}}_{x_d}^{\frac{H}{2}}-\bar{\boldsymbol{R}}_{c_d}\bar{\boldsymbol{R}}_{x_d}^{-\frac{1}{2}})^H & \boldsymbol{T}_2 \end{bmatrix}\succeq 0,$$

$$\begin{bmatrix} \sigma_v^2\boldsymbol{I}+\sigma_d^2\boldsymbol{R}_H & \boldsymbol{R}_{vd}^{\frac{1}{2}} \\ \boldsymbol{R}_{vd}^{\frac{1}{2}} & \boldsymbol{T}_3 \end{bmatrix}\succeq 0,\quad \begin{bmatrix} \boldsymbol{I} & \boldsymbol{H}^H \\ \boldsymbol{H} & \boldsymbol{R}_H \end{bmatrix}\succeq 0,$$

$$\boldsymbol{T}_i\succeq 0,\ \boldsymbol{T}_i=\boldsymbol{T}_i^H,\ i=1,2,3. \tag{50}$$

For the case of the first iteration, we have $\bar{\boldsymbol{X}}_d=\boldsymbol{0}$ and the SDP problem (50) is therefore reduced to

$$\min_{\boldsymbol{H},\boldsymbol{T}_1,\boldsymbol{T}_3,\boldsymbol{R}_H} \{N_t\text{tr}(\boldsymbol{T}_1)+N_d[\text{tr}(\boldsymbol{R}_0\boldsymbol{R}_H)+\text{tr}(\boldsymbol{T}_3)]\}$$

subject to

$$\begin{bmatrix} \sigma_v^2\boldsymbol{I}_W & (\boldsymbol{H}\boldsymbol{R}_{x_t}^{\frac{H}{2}}-\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}})^H \\ \boldsymbol{H}\boldsymbol{R}_{x_t}^{\frac{H}{2}}-\boldsymbol{R}_{c_t}\boldsymbol{R}_{x_t}^{-\frac{1}{2}} & \boldsymbol{T}_1 \end{bmatrix}\succeq 0,$$

$$\begin{bmatrix} \sigma_v^2\boldsymbol{I}+\boldsymbol{R}_H & \boldsymbol{R}_{y_d}^{\frac{1}{2}} \\ \boldsymbol{R}_{y_d}^{\frac{1}{2}} & \boldsymbol{T}_3 \end{bmatrix}\succeq 0,\quad \begin{bmatrix} \boldsymbol{I} & \boldsymbol{H}^H \\ \boldsymbol{H} & \boldsymbol{R}_H \end{bmatrix}\succeq 0,$$

$$\boldsymbol{T}_i\succeq 0,\ \boldsymbol{T}_i=\boldsymbol{T}_i^H,\ i=1,3. \tag{51}$$

The SDP problems (50) and (51) can be efficiently solved using any of the available modern interior point methods, such as *SeDuMi* [24]. The channel estimate obtained by solving (50) and (51) is denoted by $\hat{\boldsymbol{H}}_{s3}$.

C. Refining the Soft-Based ML Channel Estimate by Space-Time Signal Subspace Projection

The above soft-based ML channel estimates can be further refined by the space-time signal subspace projection [the subscript "s*i*st" stands for this projection]. By substituting the estimate $\hat{\boldsymbol{H}}_{si}$, $i=1,2,3$, in (30), we obtain $\tilde{\mathcal{H}}_{si}$ and then $\tilde{\boldsymbol{H}}_{sist}$ from (31). Notice that the pre-whitening matrix is $\mathcal{Q}_N^{\frac{1}{2}}=\frac{1}{\sigma_v}\boldsymbol{I}_{NM}$ and the rank of $\boldsymbol{P}_{\tilde{\mathcal{A}}}$ is considered as $r_{\mathcal{H}}=W+N-1$. In the case when the rank is underestimated, the projection may cause larger distortion than noise suppression even though the quality of the original channel estimate is good.

On the other hand, when the variation of the delay pattern vector $\boldsymbol{\tau}$ over L slots is smaller than the temporal resolution, the channel estimate can be further refined by substituting $\tilde{\boldsymbol{H}}_{sist}$, $i=1,2,3$, into (34). We denote the method as "s*i*stt" and the corresponding channel estimate as $\tilde{\boldsymbol{H}}_{sistt}$, $i=1,2,3$. The temporal subspace projection matrix $\boldsymbol{P}_{\tilde{U}_t}$ can be reconstructed by refined channel estimate in each iteration. According to the steps mentioned in Appendix B, $\boldsymbol{P}_{\tilde{U}_t}$ is constructed from

$$\boldsymbol{R}_t(L)=\tilde{\boldsymbol{H}}_L^H\tilde{\boldsymbol{H}}_L=\sum_{l=1}^{L}\tilde{\boldsymbol{H}}^H(l)\tilde{\boldsymbol{H}}(l), \tag{52}$$

where by the statistical property associated with $\tilde{\boldsymbol{H}}(l)$, $\tilde{\boldsymbol{H}}(l)=\hat{\boldsymbol{H}}_{s1}(l)\boldsymbol{R}_{td}^{\frac{H}{2}}(l)$ for $\hat{\boldsymbol{H}}_{s1}(l)$ and $\tilde{\boldsymbol{H}}(l)=\tilde{\boldsymbol{H}}_{s2}(l)$ for $\hat{\boldsymbol{H}}_{s2}(l)$ with $\tilde{\boldsymbol{h}}_{s2}(l)\triangleq\text{vec}(\tilde{\boldsymbol{H}}_{s2})$ and

$$\tilde{\boldsymbol{h}}_{s2}(l)=(N_t\boldsymbol{R}_{x_t}^T\otimes\boldsymbol{I}+N_d\bar{\boldsymbol{R}}_{x_d}^T\otimes\boldsymbol{G})^{\frac{H}{2}}\hat{\boldsymbol{h}}_{s2}(l). \tag{53}$$

Because the statistical property of $\hat{\boldsymbol{H}}_{s3}(l)$ is not known, the channel estimate $\hat{\boldsymbol{H}}_{s3}(l)$ can only be refined by the temporal subspace projection related to the matrix of (52) for $\hat{\boldsymbol{H}}_{s1}(l)$ or $\hat{\boldsymbol{H}}_{s2}(l)$.

V. Cramer-Rao Bounds

A. Cramer-Rao Bound of SIMO Channel Estimation With Training and Unknown Data Symbols

The mean square error (MSE) of channel estimation is lower bounded by the Cramer-Rao bound (CRB). We here derive the CRB for SIMO channel estimation with both the known training symbols and unknown data symbols. We assume that $\boldsymbol{H}=\boldsymbol{A}\boldsymbol{B}^H$ and $r_H\le\min\{M,W\}$ with $\boldsymbol{A}$ denoting a $M\times r_H$ matrix and $\boldsymbol{B}^H$ denoting a $r_H\times W$ matrix.

The corresponding CRB is given by

$$\text{CRB}(\boldsymbol{h})=\text{tr}\left[\boldsymbol{J}_c^{-1}\tilde{\boldsymbol{D}}(\tilde{\boldsymbol{D}}^H\tilde{\boldsymbol{D}})^\dagger\tilde{\boldsymbol{D}}^H\right]=\text{tr}\left(\boldsymbol{J}_c^{-1}\boldsymbol{P}_{\tilde{\boldsymbol{D}}}\right), \tag{54}$$

where $\boldsymbol{h}=\text{vec}(\boldsymbol{H})$, $\boldsymbol{P}_{\tilde{\boldsymbol{D}}}=\tilde{\boldsymbol{D}}\tilde{\boldsymbol{D}}^\dagger$, $\tilde{\boldsymbol{D}}=\boldsymbol{J}_c^{\frac{1}{2}}\boldsymbol{D}$, $\boldsymbol{D}=[\boldsymbol{B}^*\otimes\boldsymbol{I}_M,\boldsymbol{I}_W\otimes\boldsymbol{A}]$, $\boldsymbol{J}_c=N_s\boldsymbol{R}_x^T\otimes\boldsymbol{Q}^{-1}-(\boldsymbol{X}_d^*\otimes\boldsymbol{Q}^{-\frac{1}{2}})\boldsymbol{P}_{\check{\mathcal{H}}_q}(\boldsymbol{X}_d^T\otimes\boldsymbol{Q}^{-\frac{1}{2}})$, $\boldsymbol{R}_x=\frac{1}{N_s}(N_t\boldsymbol{R}_{x_t}+N_d\boldsymbol{R}_{x_d})$, $\boldsymbol{R}_{x_d}=\frac{1}{N_d}\boldsymbol{X}_d\boldsymbol{X}_d^H$, $\check{\mathcal{H}}_q=(\boldsymbol{I}_{N_d}\otimes\boldsymbol{Q}^{-\frac{1}{2}})\check{\mathcal{H}}$, and

$$\check{\mathcal{H}}=\begin{bmatrix} \boldsymbol{h}_1^H & \cdots & \boldsymbol{h}_W^H & 0 & \cdots & 0 \\ 0 & \boldsymbol{h}_1^H & \cdots & \boldsymbol{h}_W^H & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \boldsymbol{h}_1^H & \cdots & \boldsymbol{h}_W^H \end{bmatrix}^H. \tag{55}$$

The derivation of (54) is given in Appendix C.

B. Cramer-Rao Bound of Multislot SIMO Channel Estimation With Training and Unknown Data Symbols

For the multislot case, we derive the CRB of the SIMO channel estimation for $L\to\infty$. According to the channel model of multislot channel estimation (32), the CRB normalized by L, for $L\to\infty$, is given by

$$\text{CRB}(\boldsymbol{h})=\text{tr}\left\{(\boldsymbol{P}_{U_t}\otimes\boldsymbol{P}_{U_s})\mathbb{E}[\boldsymbol{J}_c^{-1}(l)]\right\}, \tag{56}$$

where $\mathbb{E}[\boldsymbol{J}_c^{-1}(l)]=\frac{1}{L}\sum_{l=1}^{L}\boldsymbol{J}_c^{-1}(l)$, $\boldsymbol{J}_c(l)=N_s\boldsymbol{R}_x^T(l)\otimes\boldsymbol{Q}^{-1}-(\boldsymbol{X}_d^*(l)\otimes\boldsymbol{Q}^{-\frac{1}{2}})\boldsymbol{P}_{\check{\mathcal{H}}_q(l)}(\boldsymbol{X}_d^T(l)\otimes\boldsymbol{Q}^{-\frac{1}{2}})$.

The derivation of (56) is given in Appendix D.

In addition, it can be easily derived that when $N_d=0$, *i.e.*, user's data is not employed in channel estimation, (56) reduces to the one as shown in [6].

C. Approximated Cramer-Rao Bound of Channel Estimation With Training and Soft Information

For the case where soft symbols are available, we derive the CRB for the joint $\{x_d(t)\}_{t=1}^{\tilde{N}_d}$ and $\boldsymbol{H}$ estimation.

By assuming that $\Delta\tilde{\boldsymbol{x}}_d = [\Delta x_d(1), \cdots, \Delta x_d(\tilde{N}_d)]$ obeys the Gaussian distribution with zeros-mean and variance $\sigma_d^2\boldsymbol{I}$ and denoting $\bar{\boldsymbol{x}}_d = [\bar{x}_d(1), \cdots, \bar{x}_d(\tilde{N}_d)]$, the log-likelihood function for the joint channel and user data symbol estimation can be expressed as

$$L_s(\boldsymbol{H}, \tilde{\boldsymbol{x}}_d) = -\frac{1}{\sigma_d^2}\|\tilde{\boldsymbol{x}}_d - \bar{\boldsymbol{x}}_d\|_F^2 - \frac{1}{\sigma_v^2}\|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{X}\|_F^2. \quad (57)$$

For simplicity, denote $L_s(\boldsymbol{H}, \tilde{\boldsymbol{x}}_d)$ as $L_s(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}$ as any estimator of $\tilde{\boldsymbol{\beta}}$, where $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ are the same as those defined in (90) and (91). According to [28], we can have that

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) - \boldsymbol{J}_{\tilde{\boldsymbol{\beta}}}^{-1} \succeq 0, \quad (58)$$

where

$$\boldsymbol{J}_{\tilde{\boldsymbol{\beta}}} = \mathbb{E}\left[\left(\frac{\partial L_s(\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{\beta}}^T}\right)^H \left(\frac{\partial L_s(\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{\beta}}^T}\right)\right] = \begin{bmatrix} \boldsymbol{J}_{\boldsymbol{\beta}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_{\boldsymbol{\beta}}^T \end{bmatrix}, \quad (59)$$

and

$$\boldsymbol{J}_{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{D}^H(N_s\boldsymbol{R}_x^T \otimes \frac{1}{\sigma_v^2}\boldsymbol{I}_M)\boldsymbol{D} & \boldsymbol{D}^H(\boldsymbol{X}_d^* \otimes \frac{1}{\sigma_v^2}\boldsymbol{I}_M)\check{\mathcal{H}} \\ \check{\mathcal{H}}^H(\boldsymbol{X}_d^T \otimes \frac{1}{\sigma_v^2}\boldsymbol{I}_M)\boldsymbol{D} & \frac{1}{\sigma_v^2}\check{\mathcal{H}}^H\check{\mathcal{H}} + \frac{1}{\sigma_d^2}\boldsymbol{I}_{\tilde{N}_d} \end{bmatrix}. \quad (60)$$

By following the same steps from (94) to (98), we can obtain the CRB for the channel estimation using soft information by

$$\mathrm{CRB}(\boldsymbol{h}) = \mathrm{tr}\left(\boldsymbol{J}_c^{-1}\boldsymbol{P}_{\tilde{\boldsymbol{D}}}\right), \quad (61)$$

where $\boldsymbol{J}_c = N_s\boldsymbol{R}_x^T \otimes \frac{1}{\sigma_v^2}\boldsymbol{I}_M - (\boldsymbol{X}_d^* \otimes \frac{1}{\sigma_v^2}\boldsymbol{I}_M)\boldsymbol{P}_{\mathcal{H}}(\boldsymbol{X}_d^T \otimes \frac{1}{\sigma_v^2}\boldsymbol{I}_M)$ and $\boldsymbol{P}_{\mathcal{H}} = \check{\mathcal{H}}(\frac{1}{\sigma_v^2}\check{\mathcal{H}}^H\check{\mathcal{H}} + \frac{1}{\sigma_d^2}\boldsymbol{I}_{\tilde{N}_d})^{-1}\check{\mathcal{H}}^H$.

VI. Numerical Results

We here conduct several numerical experiments to state the performance of the proposed channel estimation methods. To compare these methods with the existing ones, we use mean square error (MSE) and bit error rate (BER) as the performance metrics. In all the experiments, we consider that a uniform linear array with half-wavelength spacing is employed at the base station, and assume that the space-time wireless channel varies slot by slot but stays constant within each slot. We also assume Gray mapped QPSK symbols for training sequence as well as unknown data sequence. The pulse shaping function $g(t)$ is assumed to be the raised cosine impulse response with the roll-off factor of 0.22.

In the first experiment, we consider the uplink SIMO channel estimation of a wireless communication system with a single user and a base station, where the base station antenna array has four elements, *i.e.*, $M = 4$ and the received signal samples within one slot ($L = 1$) are used for the estimation. The frame structure is chosen according to the UMTS-TDD standard [22], where the length of training sequence, user data sequence, and guard interval are set to $\tilde{N}_t = 456$, $\tilde{N}_d = 1952$, and $N_g = 96$, respectively. The wireless channel is generated following the stochastic model for the bad urban case used in [5] (there are two independent clusters and each cluster has ten resolvable multipath rays) and the maximum channel length is set to $W_{max} = 57T_s$ under the UMTS-TDD standard [22]. This means that the rank of $\boldsymbol{H}$ is $r_0 = M$. In the experiment, the spatially and temporally white Gaussian noise (STWGN) is considered. 5000 channel realizations are performed for each case and the MSE of channel estimation is evaluated by $\mathbb{E}[\|\hat{\boldsymbol{H}}^{(k)} - \boldsymbol{H}\|_F^2]/\mathbb{E}[\|\boldsymbol{H}\|_F^2]$.

In Fig. 2, we plot the MSE and BER curves obtained by various channel estimation methods versus the input per-antenna-element SNR (or element SNR). These channel estimation methods include ULS by (23), RR by [5], the signal subspace projection method by (19), our proposed space-time signal subspace projection method by (31), which in Fig. 2 are abbreviated by ULS, RR, SSP, and ST, respectively. These abbreviations will be used in the subsequent figures. The dimension of the associated space-time signal subspace is determined by *Theorem 1*, where N is chosen as $\lfloor\frac{W-1}{M-1}\rfloor + 1$ with $W = W_{max}$ for ST, or $W = 30$ for lower complexity. The case of $W = 30$ is abbreviated as ST2.

From Fig. 2, we can see that the MSE of our proposed ST method outperforms ULS method by a factor of 4.5dB. It can be also seen from Fig. 2 that at high SNR, the MSE of the RR and SSP method are larger than that of the ULS method. This is because the channel rank is underestimate by either AIC or MDL information criterion when $r_0 = M$. Furthermore, the effect of distortion caused by the rank estimation error is larger than noise suppression when the SNR is high (see the analysis in [5]). Fig. 2 shows that the BER performance with RR and SSP are almost the same and slightly worse than that with ST. The BER improvements with ST over ULS and RR are about 0.7dB and 0.2dB, respectively, at BER $= 10^{-5}$.

In the second experiment, the number of antenna elements, the channel length, and the length of user data sequence are set by $M = 8$, $W = W_{max}$, and $\tilde{N}_d = 1952$, respectively, while the length of training sequence ranges from 114 to 456, and the rest parameters are the same as that in the first experiment. In this experiment we compare the MSE and BER performances of all the abovementioned estimation methods with different length of training sequence. The SNR is set to 20dB and 5000 channel realizations are performed to evaluate the performance. The simulation results are plotted in Fig. 3.

From Fig. 3, it is seen that the SSP and the ST method can achieve almost the same BER performance of the ULS with shorter length of the training sequence, and the BER of the ST method with 270 training symbols almost reaches that of the ULS with 456 training symbols. This means that about 40% training symbols can be saved without loss of BER performance.

In the third experiment, the signal samples within two slots ($L = 2$) are used for channel estimation, where the variation of the angle/delay pattern over $L = 2$ slots are assumed constant, the total length of training plus data sequences within one slot is set to $\tilde{N}_t + \tilde{N}_d = 2408$, and the rest parameters are the same as that in the first experiment. Additionally, we consider the length of training sequence within each slot as a parameter, which has two values $\tilde{N}_t = 114$ and

$\tilde{N}_t = 228$. Furthermore, we assume that the array noise is spatially correlated and Gaussian distributed (SCGN) [in the presence of unknown interference] with the covariance $\boldsymbol{Q}_{m,l} = \sigma_v^2 0.9^{|l-m|} \exp[-j\pi(l-m)\sin\theta]$ and $\theta = \pi/3$. In the experiment, we compare the performance of ULS by (23), MS [6], ST by (31), and the signal subspace projection method for multiple slots by (34). For simplicity, the last method is abbreviated by MSTT. In the experiment, 8000 channel realizations are used for each case.

The MSE curves of these methods are plotted in Fig. 4a. From the figure, we see that when $\tilde{N}_t = 114$ ST and MSTT outperforms ULS by about 9.5dB and 10dB, respectively, while when $\tilde{N}_t = 228$ ST and MSTT outperform ULS by more than 8dB and 9dB, respectively. However, the MSE of the MS method decreases slowly as SNR becomes larger and merges into that of the ULS method. This phenomena occurs in the same manner as that in the first experiment. Recall that the channel estimation accuracy with the the methods by temporal subspace projection, such as MS, depends on the length of training sequence. This means that when the number L of the slots increases, more additional training sequences can be used to estimate the associated temporal subspace. However, for the cases of smaller L, *e.g.*, $L = 1, 2$, as in the second and the third experiments, the temporal signal subspace can not be correctly estimated, such that the corresponding projection will deteriorate the initial channel estimate.

The corresponding BER curves are plotted in Fig. 4b. The black solid line is considered as the lower BER bound corresponding to the perfectly known $\boldsymbol{Q}$ and $\boldsymbol{H}$. It is seen from the figure that ST outperforms ULS by about 1.3dB and 0.5dB at BER $= 10^{-5}$ with $\tilde{N}_t = 114$ and $\tilde{N}_t = 228$, respectively. However, all the other methods perform worse than the ULS method at BER $= 10^{-5}$. This is caused by the deteriorated channel estimates from inaccurate temporal subspace projection.

In the fourth experiment, the length of the training and user data sequences are set to $\tilde{N}_t = 114$ and $\tilde{N}_d = 972$, respectively, while L ranges from 1 to 15. The rest of the parameters are the same as that in the first experiment. Also the array noise is assumed to be spatially correlated Gaussian (SCG) distributed. In the experiment, we compare the MSE and BER performance of the MS method [6], the ST method by (31), the MSTT method by (34), and the signal subspace projection method based on multiple slots by (33). The last one is abbreviated as MST. The element SNR is set to 20dB. 8000 channel realizations are performed to evaluate the performance. The performance curves with these methods are plotted in Fig. 5.

From Fig. 5, we can see that the MSE and the BER of the multislot based methods decrease by increasing L, and ST outperforms MS and MST in BER when L is small ($L \leq 2$). The MST method performs slightly better than the MS method, and the MSTT method outperforms the MS method with the same L. This means that the ST and MSTT methods are more appropriate to fast fading channels comparing to the MS method.

In the next two (fifth and sixth) experiments, the length of the training sequence, the user's data sequence, and the guard interval are set to $\tilde{N}_t = 92$, $\tilde{N}_d = 400$, and $N_g = 39$, respectively, in each slot. The number of the antenna array elements is chosen as $M = 4$. The covariance matrix of the spatially correlated Gaussian noise is assumed to be the same as that in the second experiment. The block multipath fading channels for both rank deficient ($r_H < M$) and full row rank ($r_H = M$) cases are generated according to the parameter setting of the following two cases, respectively.

Case 1 ($r_H < M$): There are $P = 6$ path rays with the direction of arrivals $\boldsymbol{\theta} = [\frac{1}{18}, \frac{1}{18}, \frac{1}{18}, \frac{1}{18}, -\frac{1}{6}, -\frac{1}{6}]\pi$, the mean path power $\boldsymbol{R}_{\boldsymbol{\alpha}} = 0.39 \times \text{diag}\{[1, 0.63, 0.40, 0.2, 0.25, 0.06]\}$, the relative path delays $\tau = [0, 0.2, 0.4, 0.6, 15, 17.2]T_s$, and the maximum temporal channel length $W_{max} = 23T_s$.

Case 2 ($r_H = M$): There are $P = 9$ resolvable path rays with the directions of arrivals $\boldsymbol{\theta} = [-40^\circ, -38^\circ, -35^\circ, -8^\circ, -4^\circ, -2^\circ, 18^\circ, 20^\circ, 25^\circ]$, the mean path power $\boldsymbol{R}_{\boldsymbol{\alpha}} = \text{diag}\{\frac{4}{41}[1, \frac{1}{2}, \frac{1}{4}, 3, \frac{3}{2}, \frac{5}{4}, 2, \frac{1}{2}, \frac{1}{4}]\}$, the relative path delays $\tau = [0, 1.2, 2.2, 5.8, 6.2, 7.2, 10.2, 11.2, 12.6]T_s$, and the maximum temporal channel length $W_{max} = 23T_s$.

To compare the average performance of channel estimation with the associated CRB, we generate 50 channel realizations with randomly varying complex path gains for each case of the parameter settings, and perform 100 Monte Carlo runs for each realization.

In the fifth experiment, we plot the CRB and the MSE curves of two channel estimation methods versus SNR for the cases of using single slot data samples. The two corresponding methods are SSP by (19) and ST by (31). The CRB is calculated according to (54). In Fig. 6 and Fig. 7, the average MSE and CRB for the estimation with these 50 channel realizations versus SNR are plotted under the channel parameters' setting of *Case* 1 and *Case* 2 .

From these two figures, we can see that the SSP and ST methods achieve almost the same MSE in the low rank case (*Case* 1) and the gap between the CRB and MSE curves is roughly 6dB with the both SSP and ST methods. However, in the full rank case (*Case* 2), the gap for the ST method is reduced to about 1dB, while the gap for the SSP method becomes larger as the SNR increases. The performance comparison with the SSP method shows that the ST method is robust to the full rank case.

In the sixth experiment, the channel parameter settings are the same as that in the fifth experiment, the element SNR is set to 20dB, L varies from 1 to 60, and all the signal samples are used. We plot in Fig. 8 the MSE and CRB curves of the channel estimation methods versus L, where the CRB is calculated by (56) using multiple slots.

Fig. 8a demonstrates for the low rank case, *i.e.*, $r_H < M$, the MSE's of the MST, MS, and MSTT methods, where the former two approach the CRB closely as L increases, while the last one suffers from the plateau. This is due to the projection matrix $\boldsymbol{P}_{\tilde{\mathcal{A}}}(l)$ in (34), which is estimated with only one slot-based signal samples. For the full row rank case, Fig. 8b shows that increasing L is not helpful to improve the performance of the MST and MS methods since the information criterion for rank estimation does not estimate the rank accurately with full rank matrix.

In all previous experiments, we didn't consider the soft-

symbols constructed by using the decoder feedback. In the following three experiments, we evaluate the effect of the soft information on the performance of the channel estimation methods. We assume that a frame of 8000 random binary equiprobable information symbols are coded by a four-state convolutional code with generators $(7, 5)$ and then permuted by a random interleaver. The coded bits are mapped on to 8000 QPSK symbols and then arranged to $L = 20$ slots. The length of the training sequence is set to $\tilde{N}_t = 92$. We choose $N = \lfloor \frac{W_{max}-1}{M-1} \rfloor + 5$ to specify the noise subspace with larger dimension. We here only consider the parameters' setting of *Case* 2 in the simulation.

In the seventh experiment, we consider the effect of the mutual information between the coded bits $b(t)$ and the corresponding LLR sequence $\lambda_1[b(t)]$ on the performance of channel estimation. According to [30], the relationship between $b(t)$ and $\lambda_1[b(t)]$ is represented by $\lambda_1[b(t)] = \frac{\sigma^2}{2} b(t) + \sigma v(t)$, where the random variable $v(t)$ obeys standard normal distribution. Then the relationship between the mutual information $\mathcal{I}$ and σ is given by [30]

$$\mathcal{I} = 1 - \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{\frac{-(\xi-\sigma^2/2)^2}{2\sigma^2}} \log_2(1 + e^{-\xi}) d\xi. \quad (62)$$

Based on these two relations, we can obtain $\lambda_1[b(t)]$ with $b(t)$ given and $\mathcal{I}$ set to a certain value of interest. By using the artificial LLR sequence as the decoder feedback in the turbo system [11], we can evaluate the effect of this mutual information on the performance of channel estimation. The MSE with the approximate ML methods by (38), (42), (51), their refinements $\tilde{\boldsymbol{H}}_{sist}, i = 1, 2, 3$, based on signal subspace projection, and the approximate CRB by (61) are plotted in Fig. 9, which are abbreviated by si, sist, $i = 1,\ 2,\ 3$, and CRB. We use $\hat{\boldsymbol{R}}_{\boldsymbol{H}} = \hat{\boldsymbol{H}}_{s1} \hat{\boldsymbol{H}}_{s1}^H$ when solving $\hat{\boldsymbol{H}}_{s2}$, and approximate the unknown variance of $\Delta x_d(t)$, *i.e.*, σ_d^2, by using a long artificial LLR sequence when calculating the CRB. The SNR is set to 3dB.

From Fig. 9, it can be seen that the performance of channel estimation is improved with the increased mutual information. When $I = 0$, the s3 method outperforms the other two methods by $\text{MSE}_{\hat{\boldsymbol{H}}_{s3}}/\text{MSE}_{\hat{\boldsymbol{H}}_{si}} \simeq 1$dB, $i = 1,\ 2$. This is because the data part $\boldsymbol{Y}_d$ make contributions in the s3 method whereas the other two methods are not able to use this data part. When $I = 1$, these three methods are equivalent since there is no approximation when the whole symbols are known. Furthermore, we can see that the performance improvements of the sist methods over the si methods are $\text{MSE}_{\hat{\boldsymbol{H}}_{si}}/\text{MSE}_{\hat{\boldsymbol{H}}_{sist}} \simeq 1.4$dB with $I = 0$ and 0.3dB with $I = 1$. This reduction by the ST subspace projection is due to the fact that the soft-based channel estimation methods keep improving the performance as the mutual information approaches 1.0, whereas the space-time signal subspace remains unchanged.

In the eighth experiment, we consider the impacts of approximating $\boldsymbol{R}_{\boldsymbol{H}}$ by its estimation on the channel estimates $\hat{\boldsymbol{H}}_{s1}$ and $\hat{\boldsymbol{H}}_{s2}$, respectively. The SNR is set to 3dB. The mutual information is set to 0.6. Since replacing $\boldsymbol{R}_{\boldsymbol{H}}$ by $\hat{\boldsymbol{R}}_H = \hat{\boldsymbol{H}}\hat{\boldsymbol{H}}$ in (36) will make the channel estimation performance dependent on the initial channel estimate $\hat{\boldsymbol{H}}$, we plot in Fig. 10 the MSE versus the MSE of initial channel estimate. It can be seen from Fig. 10 that the MSE of the channel estimate $\hat{\boldsymbol{H}}_{s2}$ is almost constant even though the MSE of the initial channel estimate varies from 0 to 0.5. Moreover, there is little impact of the initial channel estimation error on the MSE of $\hat{\boldsymbol{H}}_{s1}$ for given SNR ρ. These imply that the two methods are not so sensitive to the error of the initial channel estimate.

In the nine experiment, we analyze the performance of channel estimation in terms of BER versus SNR with the iterative receivers described in Section IV. The channel estimates $\hat{\boldsymbol{H}}_{si}$, $\hat{\boldsymbol{H}}_{sist}$ and $\hat{\boldsymbol{H}}_{sistt}$ with $i = 3$ are considered. $\hat{\boldsymbol{H}}$ obtained in the previous iteration is used to calculate $\boldsymbol{R}_0$. In Fig. 11, we plot the BER curves at the first, the second, and the fourth iterations for the s3, the s3st, the s3stt methods, respectively, which are abbreviated by s3-j, s3st-j, and s3stt-j [j denotes the iteration number]. The corresponding MSE curves are plotted in Fig. 12.

From Fig. 11, we see that s3st receiver performs slightly better than the s3 receiver and the gap between the s3st receiver and the one with exact channel knowledge is less than 2dB at the fourth iteration at BER $= 10^{-5}$. The performance of the s3stt receiver is the best among those methods presented, and is close to the known-channel lower bound (with less than 0.3dB gap at the fourth iteration at BER $= 10^{-5}$). From Fig. 12, we can see that the performance improvement of the channel estimation methods are significant at the second iteration, and the gap between the first and the second iteration are more than $\text{MSE}_1/\text{MSE}_2 = 5.6$dB. However, after the 2nd iteration, the performance improvement is not significant.

Notice that the frequency domain decision-feedback equalization (FD-DFE) method [31], [32] is used to estimate the data sequence for the first to the fourth experiments, while the soft-cancelation minimum mean square equalization (SC-MMSE) method is used for the ninth experiment.

VII. Conclusion

In this paper, we have proposed a new channel estimation methods based on space-time signal subspace projection, which is suitable for even full row-rank channel matrix or fast changing channels. Under the assumption that the variance of the delay/angle pattern can be ignored during multiple slots, we extend the channel estimation method to employing the signal samples from multiple slots. By following the approximated maximum likelihood criterion, we have also proposed an approximate ML channel estimation method using the soft information fed back from the iterative decoder. Simulation results show that all these proposed methods are efficient and outperform the existing ones.

Acknowledgment

The authors are grateful to the anonymous reviewers for their valuable comments and suggestions, which are very helpful to improve the quality of the paper.

APPENDIX A
DERIVATION OF (30)

The log-likelihood function of (28) can be expressed as

$$L(\boldsymbol{H},\boldsymbol{Q},\boldsymbol{X}_d) = -\log|\boldsymbol{Q}_N| - \frac{1}{\tilde{N}_s}\left\{\|\mathcal{Y}_t - S_N(\boldsymbol{H}\boldsymbol{X}_t)\|^2_{\boldsymbol{Q}_N^{-1}} + \|\mathcal{Y}_d - \mathcal{H}\boldsymbol{X}_D\|^2_{\boldsymbol{Q}_N^{-1}}\right\}, \tag{63}$$

where $\boldsymbol{Q}_N = \frac{N_s}{\tilde{N}_s}\boldsymbol{I}_N \otimes \boldsymbol{Q}$ and $\tilde{N}_s = \tilde{N}_t + \tilde{N}_d$.

Using similar steps for deriving (7)-(10), we can have

$$\begin{aligned} &L(\boldsymbol{H},\boldsymbol{Q},\boldsymbol{X}_d) \\ &= -\log\Big|S_N\left(\hat{\boldsymbol{H}}\boldsymbol{X}_t - \boldsymbol{H}\boldsymbol{X}_t\right)S_N\left(\hat{\boldsymbol{H}}\boldsymbol{X}_t - \boldsymbol{H}\boldsymbol{X}_t\right)^H \\ &\quad + N_t\mathcal{Q}_t + (\mathcal{Y}_d - \mathcal{H}\boldsymbol{X}_D)(\mathcal{Y}_d - \mathcal{H}\boldsymbol{X}_D)^H\Big|, \end{aligned} \tag{64}$$

where

$$\begin{aligned} \mathcal{Q}_t =& \left[\mathcal{Y}_t - S_N(\boldsymbol{H}\boldsymbol{X}_t)\right]\left[\mathcal{Y}_t - S_N(\boldsymbol{H}\boldsymbol{X}_t)\right]^H \\ &- S_N\left(\hat{\boldsymbol{H}}\boldsymbol{X}_t - \boldsymbol{H}\boldsymbol{X}_t\right)S_N\left(\hat{\boldsymbol{H}}\boldsymbol{X}_t - \boldsymbol{H}\boldsymbol{X}_t\right)^H \end{aligned}$$

is an $NM \times NM$ dimensional block Toeplitz matrix with the (i,j)-th block

$$\begin{aligned} &\boldsymbol{Q}_{i,j} \\ =&\frac{1}{N_t}[\boldsymbol{0}_{i-1}, \boldsymbol{Y}_t - \boldsymbol{H}\boldsymbol{X}_t, \boldsymbol{0}_{N-i}][\boldsymbol{0}_{j-1}, \boldsymbol{Y}_t - \boldsymbol{H}\boldsymbol{X}_t, \boldsymbol{0}_{N-j}]^H \\ &- \frac{1}{N_t}[\boldsymbol{0}_{i-1}, (\hat{\boldsymbol{H}} - \boldsymbol{H})\boldsymbol{X}_t, \boldsymbol{0}_{N-i}][\boldsymbol{0}_{j-1}, (\hat{\boldsymbol{H}} - \boldsymbol{H})\boldsymbol{X}_t, \boldsymbol{0}_{N-j}]^H \\ =&\frac{1}{N_t}[\boldsymbol{0}_{i-1}, \boldsymbol{V}_t, \boldsymbol{0}_{N-i}][\boldsymbol{0}_{j-1}, \boldsymbol{V}_t, \boldsymbol{0}_{N-j}]^H \\ &- \frac{1}{N_t}[\boldsymbol{0}_{i-1}, \boldsymbol{V}_t\boldsymbol{P}_{x_t^H}, \boldsymbol{0}_{N-i}][\boldsymbol{0}_{j-1}, \boldsymbol{V}_t\boldsymbol{P}_{x_t^H}, \boldsymbol{0}_{N-j}]^H \\ =&\begin{cases} \boldsymbol{Q}_t & i = j \\ \frac{1}{N_t}\boldsymbol{V}_t\left(\tilde{\boldsymbol{I}}_{i-j} - \boldsymbol{P}_{x_t^H}\tilde{\boldsymbol{I}}_{i-j}\boldsymbol{P}_{x_t^H}\right)\boldsymbol{V}_t^H & i \neq j \end{cases}. \end{aligned} \tag{65}$$

In (65), $\tilde{\boldsymbol{I}}_i$ is the $N_t \times N_t$ dimensional matrix with the left lower i-th diagonal entries of 1 and other entries of 0, $\boldsymbol{P}_{x_t^H} = \boldsymbol{X}_t^H(\boldsymbol{X}_t\boldsymbol{X}_t^H)^{-1}\boldsymbol{X}_t$, $\boldsymbol{0}_i$ denotes the $M \times i$ zero matrix, and $i,j \in 1,\cdots,N$. Following the standard theory in statistics under the stated assumptions [see Theorem 2.3 in [23]], we obtain

$$\begin{cases} \frac{1}{N_t}\boldsymbol{V}_t\left(\tilde{\boldsymbol{I}}_{i-j} - \boldsymbol{P}_{x_t^H}\tilde{\boldsymbol{I}}_{i-j}\boldsymbol{P}_{x_t^H}\right)\boldsymbol{V}_t^H \to \boldsymbol{0} \\ \boldsymbol{Q}_t \to \boldsymbol{Q} \end{cases}, \\ \text{with probability 1 as } N_t \to \infty, \end{cases} \tag{66}$$

which implies that $\mathcal{Q}_t \to \boldsymbol{I}_N \otimes \boldsymbol{Q}_t$ and $\boldsymbol{Q}_t \succ 0$, with probability 1 as $N_t \to \infty$. Replacing $\mathcal{Q}_t$ by $\boldsymbol{I}_N \otimes \boldsymbol{Q}_t$ and assuming $\boldsymbol{Q}_t \succ 0$, (63) can be equivalently written as

$$\begin{aligned} &L(\boldsymbol{H},\boldsymbol{Q},\boldsymbol{X}_d) \\ &= -\log\left|\left[S_N(\hat{\boldsymbol{H}}\boldsymbol{X}_t) - \mathcal{H}\boldsymbol{X}_T\right]\left[S_N(\hat{\boldsymbol{H}}\boldsymbol{X}_t) - \mathcal{H}\boldsymbol{X}_T\right]^H\right. \\ &\quad \left. + N_t\boldsymbol{I}_N \otimes \boldsymbol{Q}_t + (\mathcal{Y}_d - \mathcal{H}\boldsymbol{X}_D)(\mathcal{Y}_d - \mathcal{H}\boldsymbol{X}_D)^H\right| \\ &= -\log|N_t\boldsymbol{I}_N \otimes \boldsymbol{Q}_t| - \log\Big|\boldsymbol{I} + (\mathcal{Y}_A - \tilde{\mathcal{A}}\mathcal{B}_A^H) \\ &\quad \cdot(\mathcal{Y}_A - \tilde{\mathcal{A}}\mathcal{B}_A^H)^H\Big| \\ &= -\log|N_t\boldsymbol{I}_N \otimes \boldsymbol{Q}_t| - \log|\boldsymbol{I} + \boldsymbol{\Phi}_N| \end{aligned} \tag{67}$$

where $\tilde{\mathcal{A}} = \frac{1}{\sqrt{N_t}}(\boldsymbol{I}_N \otimes \boldsymbol{Q}_t^{-\frac{H}{2}})\mathcal{A}$,

$$\begin{aligned} \mathcal{B}_A^H &= [\mathcal{B}^H\boldsymbol{X}_T, \mathcal{B}^H\boldsymbol{X}_D], \;\; \mathcal{Y}_A = [\tilde{\mathcal{H}}, \tilde{\mathcal{Y}}_d], \\ \tilde{\mathcal{H}} &= S_N(\frac{1}{\sqrt{N_t}}\boldsymbol{Q}_t^{-\frac{H}{2}}\hat{\boldsymbol{H}}\boldsymbol{X}_t) \\ &= S_N(\frac{1}{\sqrt{N_t}}\boldsymbol{Q}_t^{-\frac{H}{2}}\hat{\boldsymbol{H}})\boldsymbol{X}_T, \end{aligned} \tag{68}$$

$\tilde{\mathcal{Y}}_d = \frac{1}{\sqrt{N_t}}(\boldsymbol{I}_N \otimes \boldsymbol{Q}_t^{-\frac{H}{2}})\mathcal{Y}_d$, and $\boldsymbol{\Phi}_N = (\mathcal{Y}_A - \tilde{\mathcal{A}}\mathcal{B}_A^H)^H(\mathcal{Y}_A - \tilde{\mathcal{A}}\mathcal{B}_A^H)$. Using the same steps as (12)-(18), we see that the log-likelihood function (67) can be maximized when

$$\mathcal{B}_A^H = \tilde{\mathcal{A}}^\dagger\mathcal{Y}_A, \tag{69}$$

$$\begin{aligned} \boldsymbol{\Phi}_N &= \mathcal{Y}_A^H(\boldsymbol{I} - \boldsymbol{P}_{\tilde{\mathcal{A}}})\mathcal{Y}_A, \\ \boldsymbol{P}_{\tilde{\mathcal{A}}} &= \boldsymbol{U}_{\tilde{\mathcal{A}}}\boldsymbol{U}_{\tilde{\mathcal{A}}}^H, \end{aligned} \tag{70}$$

where $\tilde{\mathcal{A}}^\dagger = (\tilde{\mathcal{A}}^H\tilde{\mathcal{A}})^{-1}\tilde{\mathcal{A}}^H$, $\boldsymbol{P}_{\tilde{\mathcal{A}}} = \tilde{\mathcal{A}}\tilde{\mathcal{A}}^\dagger$ denotes the projection matrix of the column space of $\tilde{\mathcal{A}}$, $\boldsymbol{U}_{\tilde{\mathcal{A}}}$ denotes the eigenvectors corresponding to $r_{\mathcal{H}}$ leading eigenvalues of $\mathcal{Y}_A\mathcal{Y}_A^H$, and the rank $r_{\mathcal{H}}$ can be estimated using AIC [20] or MDL [19] information criterion from these eigenvalues. Substituting the equations in (68) to (69) yields

$$\begin{aligned} &\mathcal{B}^H\boldsymbol{X}_T = \tilde{\mathcal{A}}^\dagger S_N\left(\frac{1}{\sqrt{N_t}}\boldsymbol{Q}_t^{-\frac{H}{2}}\hat{\boldsymbol{H}}\right)\boldsymbol{X}_T \\ \Longrightarrow &\mathcal{B}^H = \tilde{\mathcal{A}}^\dagger S_N\left(\frac{1}{\sqrt{N_t}}\boldsymbol{Q}_t^{-\frac{H}{2}}\hat{\boldsymbol{H}}\right) \end{aligned} \tag{71}$$

when $\boldsymbol{X}_T\boldsymbol{X}_T^H$ is invertible. Substituting (71), $\mathcal{A} = \sqrt{N_t}(\boldsymbol{I}_N \otimes \boldsymbol{Q}_t^{\frac{H}{2}})\tilde{\mathcal{A}}$, and (70) in (27) yields (30).

APPENDIX B
DERIVATION OF (33)

The log-likelihood function can be denoted as

$$\begin{aligned} &L(\boldsymbol{H}(1),\boldsymbol{X}_d(1),\cdots,\boldsymbol{H}(L),\boldsymbol{X}_d(L),\boldsymbol{Q}) \\ &= -\log|\boldsymbol{Q}| - \frac{1}{LN_s}\sum_{l=1}^{L}\Big(\|\boldsymbol{Y}_t(l) - \boldsymbol{H}(l)\boldsymbol{X}_t\|^2_{\boldsymbol{Q}^{-1}} \\ &\quad + \|\boldsymbol{Y}_d(l) - \boldsymbol{H}(l)\boldsymbol{X}_d(l)\|^2_{\boldsymbol{Q}^{-1}}\Big). \end{aligned} \tag{72}$$

Denote

$$\begin{aligned} &\boldsymbol{B}^H(l) = \Gamma(l)\boldsymbol{U}_t^H, \; \boldsymbol{B}_L^H = [\boldsymbol{B}^H(1),\cdots,\boldsymbol{B}^H(L)] \\ &\boldsymbol{Y}_{tL} = [\boldsymbol{Y}_t(1),\cdots,\boldsymbol{Y}_t(L)], \; \boldsymbol{X}_{tL} = \boldsymbol{I}_L \otimes \boldsymbol{X}_t, \\ &\boldsymbol{Y}_{dL} = [\boldsymbol{Y}_d(1),\cdots,\boldsymbol{Y}_d(L)], \\ &\boldsymbol{X}_{dL} = \text{Diag}(\boldsymbol{X}_d(1),\cdots,\boldsymbol{X}_d(L)), \; \boldsymbol{K}_L = \boldsymbol{B}_L^H\boldsymbol{X}_{dL}, \end{aligned} \tag{73}$$

where "Diag" transforms its argument matrices to a block diagonal matrix. (72) can be equivalently written as

$$\begin{aligned} &L(\boldsymbol{H}(1),\boldsymbol{X}_d(1),\cdots,\boldsymbol{H}(L),\boldsymbol{X}_d(L),\boldsymbol{Q}) \\ &= -\log|\boldsymbol{Q}| - \frac{1}{LN_s}\Big(\|\boldsymbol{Y}_{tL} - \boldsymbol{U}_s\boldsymbol{B}_L^H\boldsymbol{X}_{tL}\|^2_{\boldsymbol{Q}^{-1}} \\ &\quad + \|\boldsymbol{Y}_{dL} - \boldsymbol{U}_s\boldsymbol{K}_L\|^2_{\boldsymbol{Q}^{-1}}\Big). \end{aligned} \tag{74}$$

Using the similar derivation steps presented in Section II-C, we can have the ML solution

$$\Gamma(l)\tilde{\boldsymbol{U}}_t^H = \tilde{\boldsymbol{U}}_s^\dagger \tilde{\boldsymbol{H}}(l), \tag{75}$$

$$\boldsymbol{K}_L = \tilde{\boldsymbol{U}}_s^\dagger \tilde{\boldsymbol{Y}}_{dL} \Rightarrow \boldsymbol{B}^H(l)\boldsymbol{X}_d(l) = \tilde{\boldsymbol{U}}_s^\dagger \tilde{\boldsymbol{Y}}_d(l) \tag{76}$$

$$\boldsymbol{P}_{\tilde{U}_s} = \tilde{\boldsymbol{U}}_s \tilde{\boldsymbol{U}}_s^\dagger = \boldsymbol{U}_{\tilde{U}_s}\boldsymbol{U}_{\tilde{U}_s}^H, \tag{77}$$

$$\boldsymbol{Q}_t = \frac{1}{L}\sum_{l=1}^{L}\left[\boldsymbol{R}_{y_t}(l) - \boldsymbol{R}_{c_t}(l)\boldsymbol{R}_{x_t}^{-1}\boldsymbol{R}_{c_t}^H(l)\right], \tag{78}$$

where $\tilde{\boldsymbol{U}}_t = \boldsymbol{R}_{x_t}^{\frac{1}{2}}\boldsymbol{U}_t$, $\tilde{\boldsymbol{U}}_s = \boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{U}_s$, $\tilde{\boldsymbol{Y}}_{dL} = \boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{Y}_{dL}$, $\tilde{\boldsymbol{Y}}_d(l) = \boldsymbol{Q}_t^{-\frac{H}{2}}\boldsymbol{Y}_d(l)$, $\boldsymbol{Q}_t = \frac{1}{L}\sum_{l=1}^{L}\left[\boldsymbol{R}_{y_t}(l) - \boldsymbol{R}_{c_t}(l)\boldsymbol{R}_{x_t}^{-1}\boldsymbol{R}_{c_t}^H(l)\right]$, $\boldsymbol{U}_{\tilde{U}_s}$ denotes the eigenvectors corresponding to the r_s leading eigenvalues of $\boldsymbol{Y}_{AL}\boldsymbol{Y}_{AL}^H$, $\boldsymbol{Y}_{AL} = \frac{1}{\sqrt{L}}[\tilde{\boldsymbol{H}}_L, \frac{1}{\sqrt{N_t}}\tilde{\boldsymbol{Y}}_{dL}]$, and $\tilde{\boldsymbol{H}}_L = [\tilde{\boldsymbol{H}}(1), \cdots, \tilde{\boldsymbol{H}}(L)]$.

Let

$$\begin{aligned}
&\boldsymbol{A}(l) = \boldsymbol{U}_s\Gamma(l),\ \boldsymbol{A}_N = [\boldsymbol{A}^T(1), \cdots, \boldsymbol{A}^T(L)]^T\\
&\boldsymbol{Y}_{tL} = [\boldsymbol{Y}_t^T(1), \cdots, \boldsymbol{Y}_t^T(L)]^T,\\
&\boldsymbol{Y}_{dL} = [\boldsymbol{Y}_d^T(1), \cdots, \boldsymbol{Y}_d^T(L)]^T,\\
&\boldsymbol{K}_{dL} = [\boldsymbol{X}_d^H(1)\boldsymbol{B}(1), \cdots, \boldsymbol{X}_d^H(L)\boldsymbol{B}(L)]^H.
\end{aligned} \tag{79}$$

The log-likelihood function (72) can be equivalently written as

$$\begin{aligned}
&L(\boldsymbol{H}(1), \boldsymbol{X}_d(1), \cdots, \boldsymbol{H}(L), \boldsymbol{X}_d(L), \boldsymbol{Q})\\
=&-\frac{1}{L}\log|\boldsymbol{Q}_L| - \frac{1}{LN_s}\left(\|\boldsymbol{Y}_{tL} - \boldsymbol{A}_N\boldsymbol{U}_t^H\boldsymbol{X}_t\|_{\boldsymbol{Q}_L^{-1}}^2\right.\\
&\left.+\|\boldsymbol{Y}_{dL} - (\boldsymbol{I}_L\otimes\boldsymbol{U}_s)\boldsymbol{K}_{dL}\|_{\boldsymbol{Q}_L^{-1}}^2\right).
\end{aligned} \tag{80}$$

where $\boldsymbol{Q}_L = \boldsymbol{I}_L \otimes \boldsymbol{Q}$ denotes the spatial covariance of the noise. Following the steps in Section II-C and substituting (76) in (80), we can have

$$\begin{aligned}
&L(\boldsymbol{H}(1), \boldsymbol{X}_d(1), \cdots, \boldsymbol{H}(L), \boldsymbol{X}_d(L), \boldsymbol{Q})\\
=&-\log\left|N_t\left(\boldsymbol{A}_N\tilde{\boldsymbol{U}}_t^H - \boldsymbol{R}_{c_{tL}}\boldsymbol{R}_{x_t}^{-\frac{1}{2}}\right)\left(\boldsymbol{A}_N\tilde{\boldsymbol{U}}_t^H - \boldsymbol{R}_{c_{tL}}\boldsymbol{R}_{x_t}^{-\frac{1}{2}}\right)^H\right.\\
&N_t\boldsymbol{Q}_{Lt} + \left[\boldsymbol{Y}_{dL} - (\boldsymbol{I}_L\otimes\boldsymbol{U}_s\tilde{\boldsymbol{U}}_s^\dagger)\tilde{\boldsymbol{Y}}_{dL}\right]\\
&\left.\cdot\left[\boldsymbol{Y}_{dL} - (\boldsymbol{I}_L\otimes\boldsymbol{U}_s\tilde{\boldsymbol{U}}_s^\dagger)\tilde{\boldsymbol{Y}}_{dL}\right]^H\right|\\
=&-\log|N_t\boldsymbol{I}_L\otimes\boldsymbol{Q}_t| - \log|\boldsymbol{I} + \boldsymbol{R}_T + \boldsymbol{R}_D|
\end{aligned} \tag{81}$$

where $\boldsymbol{Q}_{Lt} = \boldsymbol{I}_L\otimes\boldsymbol{Q}_t$, $\boldsymbol{R}_{c_{tL}} = \frac{1}{N_t}\boldsymbol{Y}_{tL}\boldsymbol{X}_t^H$ and $\tilde{\boldsymbol{Y}}_{dL} = (\boldsymbol{I}_L\otimes\boldsymbol{Q}_t^{-\frac{H}{2}})\tilde{\boldsymbol{Y}}_{dL}$, $\boldsymbol{Q}_t$ is the same as defined in (78),

$$\boldsymbol{R}_D = \left[\boldsymbol{I}_L\otimes(\boldsymbol{I} - \boldsymbol{P}_{\tilde{U}_s})\right]\tilde{\boldsymbol{Y}}_{dL}\tilde{\boldsymbol{Y}}_{dL}^H\left[\boldsymbol{I}_L\otimes(\boldsymbol{I} - \boldsymbol{P}_{\tilde{U}_s})\right]^H, \tag{82}$$

$$\begin{aligned}
\boldsymbol{R}_T &= \left[\tilde{\boldsymbol{H}}_L - \tilde{\boldsymbol{A}}_N\tilde{\boldsymbol{U}}_t^H\right]\left[\tilde{\boldsymbol{H}}_L - \tilde{\boldsymbol{A}}_N\tilde{\boldsymbol{U}}_t^H\right]^H\\
&= \left[\tilde{\boldsymbol{A}}_N - \tilde{\boldsymbol{H}}_L\tilde{\boldsymbol{U}}_t^{H\dagger}\right]\tilde{\boldsymbol{U}}_t^H\tilde{\boldsymbol{U}}_t\left[\tilde{\boldsymbol{A}}_N - \tilde{\boldsymbol{H}}_L\tilde{\boldsymbol{U}}_t^{H\dagger}\right]^H\\
&\quad + \tilde{\boldsymbol{H}}_L\tilde{\boldsymbol{H}}_L^H - \tilde{\boldsymbol{H}}_L\boldsymbol{P}_{\tilde{U}_t}\tilde{\boldsymbol{H}}_L^H,
\end{aligned} \tag{83}$$

$\tilde{\boldsymbol{H}}_L = [\tilde{\boldsymbol{H}}^T(1), \cdots, \tilde{\boldsymbol{H}}^T(L)]^T$, $\tilde{\boldsymbol{A}}_N = (\boldsymbol{I}_L\otimes\boldsymbol{Q}_t^{-\frac{H}{2}})\boldsymbol{A}_N$, and $\boldsymbol{P}_{\tilde{U}_t} = \tilde{\boldsymbol{U}}_t^{H\dagger}\tilde{\boldsymbol{U}}_t^H$. Let $\boldsymbol{P}_{\tilde{U}_{t0}} = \boldsymbol{U}_{\tilde{U}_t}\boldsymbol{U}_{\tilde{U}_t}^H$ and $\boldsymbol{U}_{\tilde{U}_t}$ denotes the eigenvectors corresponding to r_t leading eigenvalues of $\tilde{\boldsymbol{H}}_L^H\tilde{\boldsymbol{H}}_L$, $r_t \le \min(LM, W)$. It is evident that $\tilde{\boldsymbol{H}}_L\boldsymbol{P}_{\tilde{U}_{t0}}\tilde{\boldsymbol{H}}_L^H \succeq \tilde{\boldsymbol{H}}_L\boldsymbol{P}_{\tilde{U}_t}\tilde{\boldsymbol{H}}_L^H$ by using eigenvalue decomposition and then we have

$$\boldsymbol{I} + \boldsymbol{R}_T + \boldsymbol{R}_D \succeq \boldsymbol{I} + \boldsymbol{R}_D + \tilde{\boldsymbol{H}}_L\tilde{\boldsymbol{H}}_L^H - \tilde{\boldsymbol{H}}_L\boldsymbol{P}_{\tilde{U}_{t0}}\tilde{\boldsymbol{H}}_L^H, \tag{84}$$

where the equality holds if and only if

$$\tilde{\boldsymbol{A}}_N = \tilde{\boldsymbol{H}}_L\tilde{\boldsymbol{U}}_t^{H\dagger} \Rightarrow \tilde{\boldsymbol{U}}_s\Gamma(l) = \tilde{\boldsymbol{H}}(l)\tilde{\boldsymbol{U}}_t^{H\dagger}, \tag{85}$$

$$\boldsymbol{P}_{\tilde{U}_t} = \boldsymbol{P}_{\tilde{U}_{t0}}. \tag{86}$$

In other words, the ML solution to (72) can also be approached by (85)-(86). $\Gamma(l)$ that satisfies (75) and (85) simultaneously can be obtained by

$$\Gamma(l) = \tilde{\boldsymbol{U}}_s^\dagger\tilde{\boldsymbol{H}}(l)\tilde{\boldsymbol{U}}_t^{H\dagger}. \tag{87}$$

Substituting (87), $\boldsymbol{U}_t = \boldsymbol{R}_{x_t}^{-\frac{1}{2}}\tilde{\boldsymbol{U}}_t$, and $\boldsymbol{U}_s = \boldsymbol{Q}_t^{\frac{H}{2}}\tilde{\boldsymbol{U}}_s$ in (32) yields (33).

Appendix C
Derivation of (54)

To facilitate the derivation, we rewrite the log-likelihood function of joint channel estimation and unknown data symbol detection as

$$\begin{aligned}
&L_s(\boldsymbol{Q}, \boldsymbol{H}, \tilde{\boldsymbol{x}}_d)\\
=&-N_s\log|\boldsymbol{Q}| - \|\boldsymbol{Y}_t - \boldsymbol{H}\boldsymbol{X}_t\|_{\boldsymbol{Q}^{-1}}^2 - \|\boldsymbol{y}_d - \check{\mathcal{H}}\tilde{\boldsymbol{x}}_d\|_{\boldsymbol{I}_{N_d}\otimes\boldsymbol{Q}^{-1}}^2\\
=&-N_s\log|\boldsymbol{Q}| - \|\boldsymbol{y} - (\boldsymbol{X}^T\otimes\boldsymbol{I}_M)(\boldsymbol{I}_W\otimes\boldsymbol{A})\boldsymbol{b}\|_{\boldsymbol{I}_{N_s}\otimes\boldsymbol{Q}^{-1}}^2\\
=&-N_s\log|\boldsymbol{Q}| - \|\boldsymbol{y} - (\boldsymbol{X}^T\otimes\boldsymbol{I}_M)(\boldsymbol{B}^*\otimes\boldsymbol{I}_M)\boldsymbol{a}\|_{\boldsymbol{I}_{N_s}\otimes\boldsymbol{Q}^{-1}}^2
\end{aligned} \tag{88}$$

where $\tilde{\boldsymbol{x}}_d = [x_d(1), \cdots, x_d(\tilde{N}_d)]^T$ is the unknown data sequence, $\boldsymbol{h} = \text{vec}(\boldsymbol{H})$, $\boldsymbol{y} = \text{vec}(\boldsymbol{Y})$, $\boldsymbol{a} = \text{vec}(\boldsymbol{A})$, and $\boldsymbol{b} = \text{vec}(\boldsymbol{B}^H)$.

We define $\boldsymbol{q} = \text{vec}(\boldsymbol{Q})$, $\boldsymbol{\beta} = \left[\boldsymbol{a}^T, \boldsymbol{b}^T, \tilde{\boldsymbol{x}}_d^T\right]^T$, $\tilde{\boldsymbol{q}} = \left[\boldsymbol{q}^T, \boldsymbol{q}^H\right]^T$, $\tilde{\boldsymbol{\beta}} = \left[\boldsymbol{\beta}^T, \boldsymbol{\beta}^H\right]^T$, $\boldsymbol{r} = \left[\tilde{\boldsymbol{q}}^T, \tilde{\boldsymbol{\beta}}^T\right]^T$, and $\hat{\boldsymbol{r}}$ is any unbiased estimator of $\boldsymbol{r}$. $L_s(\boldsymbol{Q}, \boldsymbol{H}, \tilde{\boldsymbol{x}}_d)$ is replaced by $L_s(\boldsymbol{q}, \boldsymbol{\beta})$. According to [28], we have:

$$\text{cov}(\hat{\boldsymbol{r}}) - \boldsymbol{J}_{\boldsymbol{r}}^{-1} \succeq 0, \tag{89}$$

where "cov" is the notation to denote covariance matrix and

$$\boldsymbol{J}_{\boldsymbol{r}} = \mathbb{E}\left[\left(\frac{\partial L_s(\boldsymbol{q}, \boldsymbol{\beta})}{\partial \boldsymbol{r}^T}\right)^H\left(\frac{\partial L_s(\boldsymbol{q}, \boldsymbol{\beta})}{\partial \boldsymbol{r}^T}\right)\right]. \tag{90}$$

Define

$$\boldsymbol{J}_{\tilde{\boldsymbol{q}}} = \mathbb{E}\left[\left(\frac{\partial L_s(\boldsymbol{q},\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{q}}^T}\right)^H \left(\frac{\partial L_s(\boldsymbol{q},\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{q}}^T}\right)\right],$$
$$\boldsymbol{J}_{\tilde{\boldsymbol{q}}\tilde{\boldsymbol{\beta}}} = \mathbb{E}\left[\left(\frac{\partial L_s(\boldsymbol{q},\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{q}}^T}\right)^H \left(\frac{\partial L_s(\boldsymbol{q},\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{\beta}}^T}\right)\right],$$
$$\boldsymbol{J}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{q}}} = \mathbb{E}\left[\left(\frac{\partial L_s(\boldsymbol{q},\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{\beta}}^T}\right)^H \left(\frac{\partial L_s(\boldsymbol{q},\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{q}}^T}\right)\right],$$
$$\boldsymbol{J}_{\tilde{\boldsymbol{\beta}}} = \mathbb{E}\left[\left(\frac{\partial L_s(\boldsymbol{q},\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{\beta}}^T}\right)^H \left(\frac{\partial L_s(\boldsymbol{q},\boldsymbol{\beta})}{\partial \tilde{\boldsymbol{\beta}}^T}\right)\right]. \quad (91)$$

We have $\boldsymbol{J}_{\tilde{\boldsymbol{q}}\tilde{\boldsymbol{\beta}}} = \boldsymbol{0}$, $\boldsymbol{J}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{q}}} = \boldsymbol{0}$, and

$$\boldsymbol{J}_{\tilde{\boldsymbol{\beta}}} = \begin{bmatrix} \boldsymbol{J}_{\boldsymbol{\beta}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_{\boldsymbol{\beta}}^T \end{bmatrix},$$
$$\boldsymbol{J}_{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{D}^H(N_s\boldsymbol{R}_x^T \otimes \boldsymbol{Q}^{-1})\boldsymbol{D} & \boldsymbol{D}^H(\boldsymbol{X}_d^* \otimes \boldsymbol{Q}^{-1})\check{\mathcal{H}} \\ \check{\mathcal{H}}^H(\boldsymbol{X}_d^T \otimes \boldsymbol{Q}^{-H})\boldsymbol{D} & \check{\mathcal{H}}_q^H\check{\mathcal{H}}_q \end{bmatrix}$$
$$= \mathcal{D}^H \begin{bmatrix} N_s\boldsymbol{R}_x^T \otimes \boldsymbol{Q}^{-1} & (\boldsymbol{X}_d^* \otimes \boldsymbol{Q}^{-1})\check{\mathcal{H}} \\ \check{\mathcal{H}}^H(\boldsymbol{X}_d^T \otimes \boldsymbol{Q}^{-H}) & \check{\mathcal{H}}_q^H\check{\mathcal{H}}_q \end{bmatrix} \mathcal{D}, \quad (92)$$

where $\boldsymbol{D} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\beta}_1^T} = [\boldsymbol{B}^* \otimes \boldsymbol{I}_M, \boldsymbol{I}_W \otimes \boldsymbol{A}]$, $\boldsymbol{\beta}_1 = \left[\boldsymbol{a}^T, \boldsymbol{b}^T\right]^T$, $\boldsymbol{R}_x = \frac{1}{N_s}(N_t\boldsymbol{R}_{x_t} + N_d\boldsymbol{R}_{x_d})$, $\boldsymbol{R}_{x_d} = \frac{1}{N_d}\boldsymbol{X}_d\boldsymbol{X}_d^H$, $\check{\mathcal{H}}_q = (\boldsymbol{I}_{N_d} \otimes \boldsymbol{Q}^{-\frac{1}{2}})\check{\mathcal{H}}$, $(\cdot)^*$ denotes the complex conjugate and

$$\mathcal{D} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}.$$

According to the definition of (91), the associated terms of $\boldsymbol{J}_{\boldsymbol{r}}$ in (90) can be reduced, such that

$$\boldsymbol{J}_{\boldsymbol{r}} = \begin{bmatrix} \boldsymbol{J}_{\tilde{\boldsymbol{q}}} & \boldsymbol{J}_{\tilde{\boldsymbol{q}}\tilde{\boldsymbol{\beta}}} \\ \boldsymbol{J}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{q}}} & \boldsymbol{J}_{\tilde{\boldsymbol{\beta}}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{J}_{\tilde{\boldsymbol{q}}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_{\tilde{\boldsymbol{\beta}}} \end{bmatrix}. \quad (93)$$

Notice that the rank of the $MW \times (W+M)r_H$ matrix $\boldsymbol{D}$ is $(W+M-r_H)r_H$ [1], and that $\boldsymbol{J}_{\boldsymbol{\beta}}$, $\boldsymbol{J}_{\tilde{\boldsymbol{\beta}}}$, and $\boldsymbol{J}_r$ will be rank deficient when $\mathcal{D}$ is not full column rank.

Based on the results in [27] and [28], the CRB of the channel estimation can be derived as

$$\text{CRB}(\boldsymbol{h}) = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{r}^T}\boldsymbol{J}_{\boldsymbol{r}}^{\dagger}\left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{r}^T}\right)^H$$
$$= [\boldsymbol{0}, \boldsymbol{D}, \boldsymbol{0}]\boldsymbol{J}_{\boldsymbol{r}}^{\dagger}[\boldsymbol{0}, \boldsymbol{D}, \boldsymbol{0}]^H, \quad (94)$$

where $\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{r}^T} = [\boldsymbol{0}, \boldsymbol{D}, \boldsymbol{0}]$, and $\boldsymbol{J}_{\boldsymbol{r}}^{\dagger}$ is a Moore-Penrose pseudoinverse notation of $\boldsymbol{J}_{\boldsymbol{r}}$. According to the definition of Moore-Penrose pseudoinverse, $\boldsymbol{J}_{\boldsymbol{r}}^{\dagger}$ can be denoted by

$$\boldsymbol{J}_{\boldsymbol{r}}^{\dagger} = \begin{bmatrix} \boldsymbol{J}_{\tilde{\boldsymbol{q}}}^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_{\boldsymbol{\beta}}^{\dagger} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{J}_{\boldsymbol{\beta}}^{T\dagger} \end{bmatrix}. \quad (95)$$

Denoting $\boldsymbol{J}_1 = \boldsymbol{D}^H(N_s\boldsymbol{R}_x^T \otimes \boldsymbol{Q}^{-1})\boldsymbol{D}$, $\boldsymbol{J}_2 = \boldsymbol{D}^H(\boldsymbol{X}_d^* \otimes \boldsymbol{Q}^{-1})\check{\mathcal{H}}$, $\boldsymbol{J}_3 = \check{\mathcal{H}}_q^H\check{\mathcal{H}}_q$, $\boldsymbol{P}_{\check{\mathcal{H}}_q} = \check{\mathcal{H}}_q\boldsymbol{J}_3^{-1}\check{\mathcal{H}}_q^H$, $\boldsymbol{J}_c = N_s\boldsymbol{R}_x^T \otimes \boldsymbol{Q}^{-1} - (\boldsymbol{X}_d^* \otimes \boldsymbol{Q}^{-\frac{1}{2}})\boldsymbol{P}_{\check{\mathcal{H}}_q}(\boldsymbol{X}_d^T \otimes \boldsymbol{Q}^{-\frac{1}{2}})$, and $\boldsymbol{J}_4 = \boldsymbol{D}^H\boldsymbol{J}_c\boldsymbol{D}$, we have

$$\boldsymbol{J}_{\boldsymbol{\beta}}^{\dagger} = \begin{bmatrix} \boldsymbol{J}_4^{\dagger} & -\boldsymbol{J}_4^{\dagger}\boldsymbol{J}_2\boldsymbol{J}_3^{-1} \\ -(\boldsymbol{J}_4^{\dagger}\boldsymbol{J}_2\boldsymbol{J}_3^{-1})^H & \boldsymbol{J}_3^{-1} + \boldsymbol{J}_3^{-1}\boldsymbol{J}_2^H\boldsymbol{J}_4^{\dagger}\boldsymbol{J}_2\boldsymbol{J}_3^{-1} \end{bmatrix}. \quad (96)$$

Notice that (96) is derived by substituting

$$\boldsymbol{P}_{\boldsymbol{J}_4} = \boldsymbol{J}_4\boldsymbol{J}_4^{\dagger} = \tilde{\boldsymbol{D}}^H\tilde{\boldsymbol{D}}(\tilde{\boldsymbol{D}}^H\tilde{\boldsymbol{D}})^{\dagger} = \boldsymbol{P}_{\tilde{\boldsymbol{D}}^H},$$
$$\boldsymbol{P}_{\boldsymbol{J}_4}\boldsymbol{J}_1 = \boldsymbol{P}_{\tilde{\boldsymbol{D}}^H}\tilde{\boldsymbol{D}}^H\boldsymbol{J}_c^{-\frac{1}{2}}(\boldsymbol{R}_x^T \otimes \boldsymbol{Q}^{-1})\boldsymbol{D} = \boldsymbol{J}_1,$$
$$\boldsymbol{P}_{\boldsymbol{J}_4}\boldsymbol{J}_2 = \boldsymbol{J}_2, \quad (97)$$

in the associated Moore-Penrose pseudoinverse matrix, where $\tilde{\boldsymbol{D}} = \boldsymbol{J}_c^{\frac{1}{2}}\boldsymbol{D}$ and $\boldsymbol{P}_{\tilde{\boldsymbol{D}}^H} = \tilde{\boldsymbol{D}}^H\tilde{\boldsymbol{D}}^{H\dagger}$ is the projection matrix. Substituting (96) in (94), the CRB of channel estimation can be obtained by

$$\text{CRB}(\boldsymbol{h}) = \text{tr}\left(\boldsymbol{D}\boldsymbol{J}_4^{\dagger}\boldsymbol{D}^H\right). \quad (98)$$

(54) can be obtained by substituting $\boldsymbol{J}_4$ in (98).

Appendix D
Derivation of (56)

According to the channel model of multislot channel estimation (32), the unknown parameters are $\boldsymbol{\beta}_1 = \left[\boldsymbol{u}_s^T, \boldsymbol{u}_t^T, \boldsymbol{v}_1^T, \cdots, \boldsymbol{v}_L^T,\right]^T$, where $\boldsymbol{u}_s = \text{vec}(\boldsymbol{U}_s)$, $\boldsymbol{u}_t = \text{vec}(\boldsymbol{U}_t^H)$, $\boldsymbol{v}_l = \text{vec}(\Gamma(l))$, and $\boldsymbol{H}(l) = \boldsymbol{U}_s\Gamma(l)\boldsymbol{U}_t^H$. Let $\boldsymbol{h} = \text{vec}[\boldsymbol{H}(1), \cdots, \boldsymbol{H}(L)]$ and $\boldsymbol{D} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\beta}_1^T}$. According to [28] and the steps in the previous section, we have:

$$\text{CRB}(\boldsymbol{h}) = \text{tr}\left\{\boldsymbol{D}\boldsymbol{J}_4^{\dagger}\boldsymbol{D}^H\right\}, \quad (99)$$

where

$$\boldsymbol{J}_4 = \begin{bmatrix} \boldsymbol{J}_{uu} & \boldsymbol{J}_{uv} \\ \boldsymbol{J}_{uv}^H & \boldsymbol{J}_{vv} \end{bmatrix},$$
$$\boldsymbol{J}_{uu} = \sum_{l=1}^{L}\left\{\tilde{\boldsymbol{D}}_l^H\boldsymbol{R}_p(l)\tilde{\boldsymbol{D}}_l\right\},$$
$$\boldsymbol{J}_{vv} = \text{diag}\left\{\left[\tilde{\boldsymbol{D}}_{v1}^H\boldsymbol{R}_p(1)\tilde{\boldsymbol{D}}_{v1}, \cdots, \tilde{\boldsymbol{D}}_{vL}^H\boldsymbol{R}_p(L)\tilde{\boldsymbol{D}}_{vL}\right]\right\},$$
$$\boldsymbol{J}_{uv} = \left[\tilde{\boldsymbol{D}}_1^H\boldsymbol{R}_p(1)\tilde{\boldsymbol{D}}_{v1}, \cdots, \tilde{\boldsymbol{D}}_L^H\boldsymbol{R}_p(L)\tilde{\boldsymbol{D}}_{vL}\right], \quad (100)$$

$\boldsymbol{R}_p(l) = \boldsymbol{I} - \frac{1}{N_s}(\boldsymbol{R}_x^{-\frac{T}{2}}(l)\boldsymbol{X}_d^*(l) \otimes \boldsymbol{I})\boldsymbol{P}_{\check{\mathcal{H}}_q(l)}(\boldsymbol{X}_d^T(l)\boldsymbol{R}_x^{-\frac{*}{2}}(l) \otimes \boldsymbol{I})$, $\tilde{\boldsymbol{D}}_l = (\sqrt{N_s}\boldsymbol{R}_x^{\frac{T}{2}}(l) \otimes \boldsymbol{Q}^{-\frac{1}{2}})[(\boldsymbol{U}_t^*\Gamma^T(l)) \otimes \boldsymbol{I}, \boldsymbol{I} \otimes (\boldsymbol{U}_s\Gamma(l))]$, and $\tilde{\boldsymbol{D}}_{vl} = (N_s\boldsymbol{R}_x^T(l) \otimes \boldsymbol{Q}^{-1})^{\frac{1}{2}}(\boldsymbol{U}_t^* \otimes \boldsymbol{U}_s)$. According to the properties for the pseudoinverse of a partitioned matrix [29], the blocks of $\boldsymbol{J}_4^{\dagger}$ corresponding to $\boldsymbol{J}_{uu}$, $\boldsymbol{J}_{uv}$, and $\boldsymbol{J}_{uv}^H$ are $\mathcal{O}(1/L)$. It follows that the channel estimation error uniquely depends on $\{\boldsymbol{v}_l\}_{l=1}^L$ as $L \to \infty$. Hence, the CRB normalized by L, for $L \to \infty$, is given by

$$\text{CRB}(\boldsymbol{h}) = \frac{1}{L}\sum_{l=1}^{L}\text{tr}\left\{(\boldsymbol{U}_t^* \otimes \boldsymbol{U}_s)(\tilde{\boldsymbol{D}}_{vl}^H\boldsymbol{R}_p(l)\tilde{\boldsymbol{D}}_{vl})^{-1} \cdot (\boldsymbol{U}_t^T \otimes \boldsymbol{U}_s^H)\right\}. \quad (101)$$

Substituting $\boldsymbol{R}_p(l)$ and $\tilde{\boldsymbol{D}}_{vl}$ in (101) yields (56).

REFERENCES

[1] P. Stoica and M. Viberg, "Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions," *IEEE Trans. Signal Process.*, vol. 44, pp. 3069-3078, 1996.

[2] M. Nicoli, "Multiuser Reduced Rank Receivers for TD/CDMA Systems," Ph.D Thesis, Politecnico di Milano, Italy, Dec.2001.

[3] D. Giancola, A. Sanguanini, and U. Spagnolini, "Variable rank receiver structures for low-rank space-time channels," in *Proc. IEEE Veh. Technol. Conf.*, 1999, pp. 65-69 vol.1.

[4] E. Lindskog and C. Tidestav, "Reduced rank channel estimation," in *Proc. IEEE Veh. Technol. Conf.*, May 1999, pp. 1126-1130 vol.2.

[5] M. Nicoli and U.Spagnolini, "Reduced-rank channel estimation for time-slotted mobile communication systems," *IEEE Trans. Signal Process.*, vol. 53, pp. 926-944, 2005.

[6] M. Nicoli, O. Simeone, and U. Spagnolini, "Multislot estimation of fast-varying space-time communication channels," *IEEE Trans. Signal Process.*, vol. 51, pp. 1184-1195, 2003.

[7] M. Nicoli, O. Simeone, and U. Spagnolini, "Multislot estimation of frequency-selective fast-varying channels," *IEEE Trans. Commun.*, vol. 51, pp. 1337-1347, 2003.

[8] R. Koetter, A. C. Singer, and M. Tuchler, "Turbo equalization," *IEEE Signal Process. Mag.*, vol.21, no.1, pp. 67-80, Jan. 2004.

[9] M. Nicoli and U. Spagnolini, "A subspace method for channel estimation in soft-iterative receivers," in *Proc. 13th Eur. Signal Process. Conf.*, Sep. 2005.

[10] S. Ferrara, T. Matsumoto, M. Nicoli, and U. Spagnolini, "Soft Iterative Channel Estimation With Subspace and Rank Tracking," *IEEE Signal Process. Lett.*, vol. 14, pp. 5-8, 2007.

[11] M. Nicoli, S. Ferrara, and U. Spagnolini, "Soft-Iterative Channel Estimation: Methods and Performance Analysis," *IEEE Trans. Signal Process.*, vol. 55, pp. 2993-3006, 2007.

[12] M. Loncar, R. Muller, J. Wehinger, C. Mecklenbrauker, and T. Abe, "Iterative channel estimation and data detection in frequency-selective fading MIMO channels," *Eur. Trans. Telecommun.*, vol. 15, no. 4, pp. 459-470, Sep. 2004.

[13] M. Nicoli and U. Spagnolini, "Subspace-methods for space-time processing," in *Smart Antennas - State of the Art*, edited by T. Kaiser, A. Bourdoux, H. Boche, J. R. Fonollosa, J.B. Andersen, W. Utschitch, Part I Receiver, EURASIP-Book series on Signal Processing and Communications, Hindawi Publ. Corp., 2005.

[14] V. Buchoux, O. Cappe, E. Moulines, and A. Gorokhov, "On the performance of semi-blind subspace-based channel estimation," *IEEE Trans. Signal Process.*, vol. 48, pp. 1750-1759, 2000.

[15] E. de Carvalho and D. Slock, "Deterministic quadratic semi-blind FIR multichannel estimation: algorithms and performance," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, Jun. 5-9, 2000, vol. 5, pp. 2553-2556.

[16] Soderstrom and P. Stoica, *System Identification*. London: Prentice-Hall, 1989.

[17] M. Viberg, P. Stoica, and B. Ottersten, "Maximum likelihood array processing in spatially correlated noise fields using parameterized signals," *IEEE Trans. Signal Process.*, vol. 45, pp. 996-1004, 1997.

[18] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge university press, 2005.

[19] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 387-392, 1985.

[20] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Inform. Theory*, B. N. Petrov and F. Caski, Eds., 1973, pp. 267-281.

[21] J. Hagenauer, "The turbo principle: Tutorial introduction and state of the art," in *Proc. Int. Symp. on Turbo Codes*, Brest, France, Sept. 1997, pp.1-11.

[22] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. New York: Wiley, 2000.

[23] L. Ljung, System Identification: Theory for the User. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[24] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," Optim. Meth. Softw., vol. 11-12, pp. 625-653, Aug. 1999.

[25] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1991.

[26] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices," in *Proc. American Control Conf.*, 2003, pp. 2156-2162 vol.3.

[27] H. L. Van Trees, *Optimum Array Processing*, Part IV of Detection, Estimation, and Modulation Theory, A John Wiley and Sons, Inc. 2002.

[28] A. van den Bos, "A Cramer-Rao lower bound for complex parameters," *IEEE Trans. Signal Process.*, vol.42, no.10, pp.2859, Oct 1994.

[29] W. E. Larimore, "Order-recursive factorization of the pseudoinverse of a covariance matrix," *IEEE Trans. Automat Contr.*, vol. 35, pp. 1299-1303, 1990.

[30] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 49, pp. 1727-1737, Oct. 2001.

[31] J. Tubbax, L. Van der Perre, S. Donnay, and M. Engels, "Single-carrier communication using decision-feedback equalization for multiple antennas," in *Proc. ICC '03*, 2003, pp. 2321-2325 vol.4.

[32] D. Falconer, S. L. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Commun. Mag.*, vol. 40, pp. 58-66, 2002.

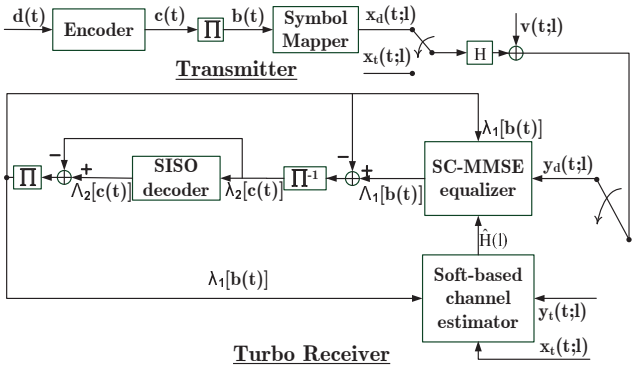
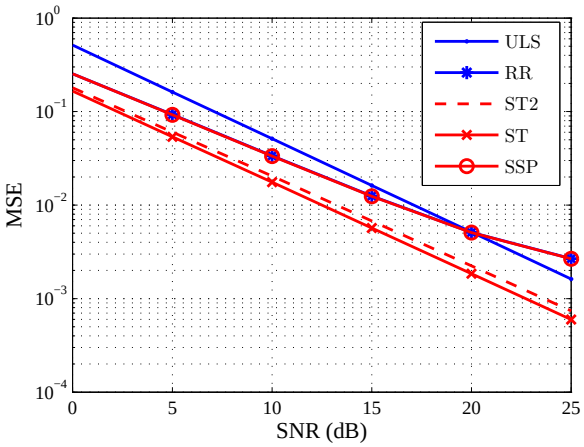
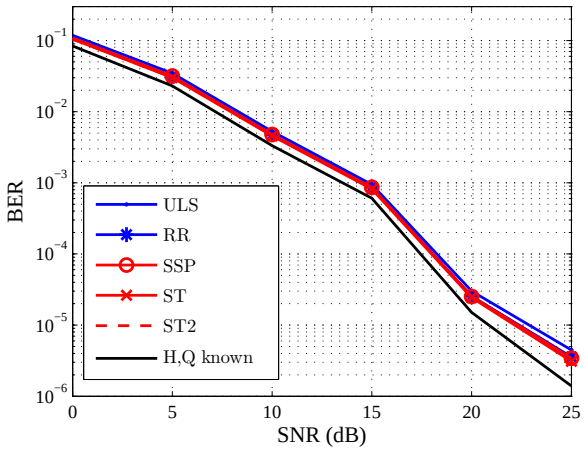


Fig. 1: The system framework: $\{d(t)\}$, $\{c(t)\}$, and $\{b(t)\}$ are sequences of information symbols, code bits, and interleaved code bits, respectively; $\boldsymbol{x}_t(t;l)$ and $\boldsymbol{y}_t(t;l)$ are the t-th training symbol in the l-th slot and the corresponding sample at the receiver; user data symbols $\boldsymbol{x}_d(t;l)$ and $\boldsymbol{y}_d(t;l)$ are similarly defined; Π and Π^{-1} denotes interleaver and deinterleaver; $\lambda_1[b(t)]$ ($\lambda_2[c(t)]$) and $\Lambda_1[b(t)]$ ($\Lambda_2[c(t)]$) denotes *a priori* LLR and *a posteriori* LLR for $b(t)$ ($c(t)$), respectively.

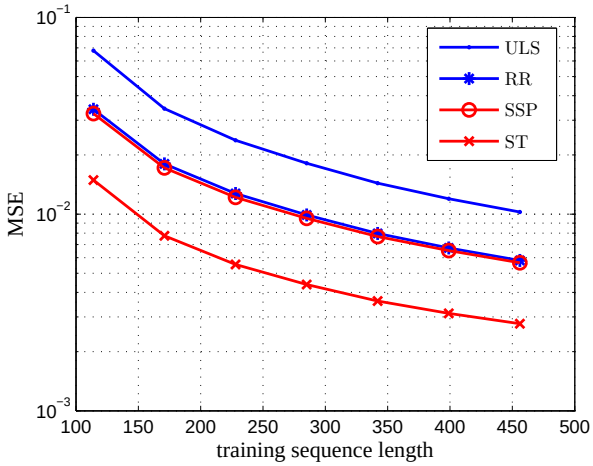


(a) MSE of channel estimation methods

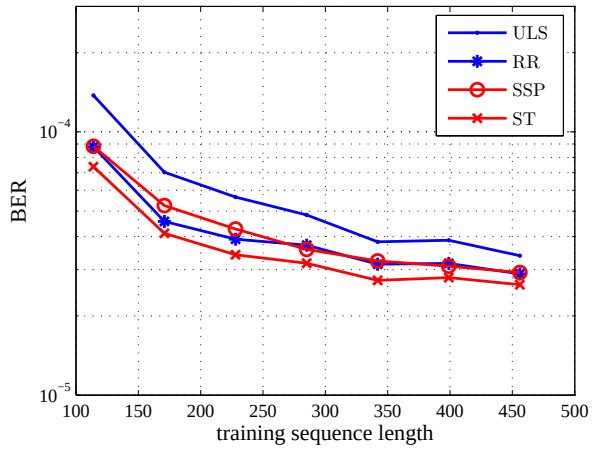


(b) BER corresponding to channel estimation methods

Fig. 2: The MSE and BER of channel estimation without soft information versus SNR for single slot 4×57 channel matrix and STWGN.

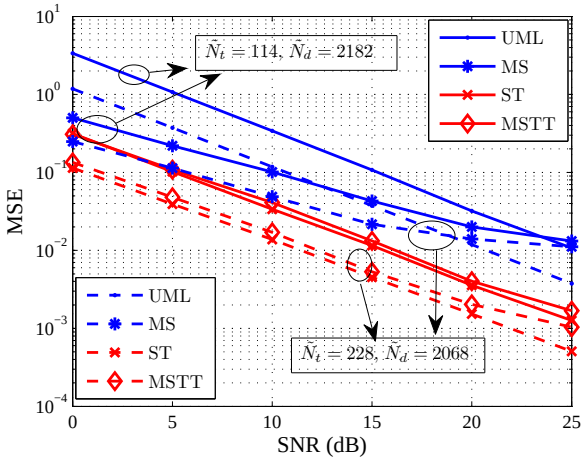


(a) MSE of channel estimation methods

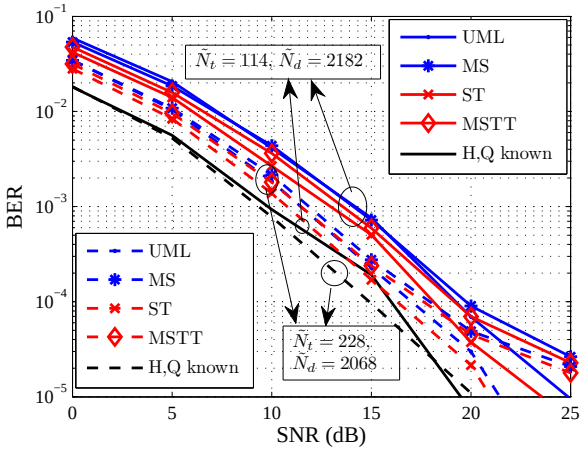


(b) BER corresponding to channel estimation methods

Fig. 3: The MSE and BER of channel estimation without soft information versus the length of training sequence for single slot 8×57 channel matrix and STWGN.

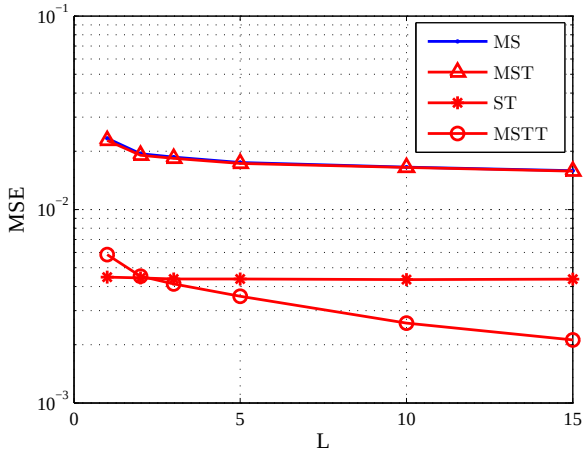


(a) MSE of channel estimation methods

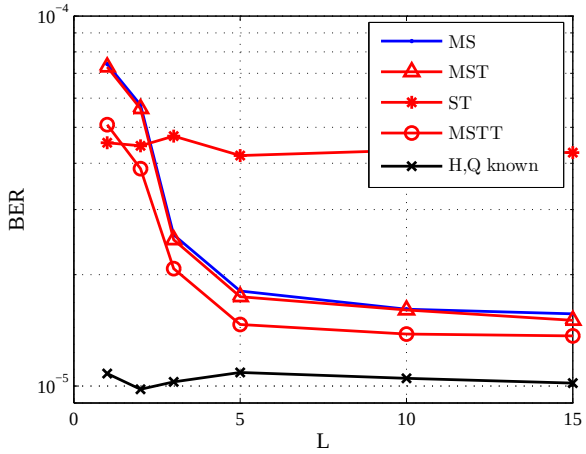


(b) BER corresponding to channel estimation methods

Fig. 4: The MSE and BER of channel estimation without soft information versus SNR for $L = 2$ slots 4×57 channel matrix and SCGN.



(a) MSE of channel estimation methods



(b) BER corresponding to channel estimation methods

Fig. 5: The MSE and BER of channel estimation without soft information versus L for $SNR = 20$dB, 4×57 channel matrix and SCGN.

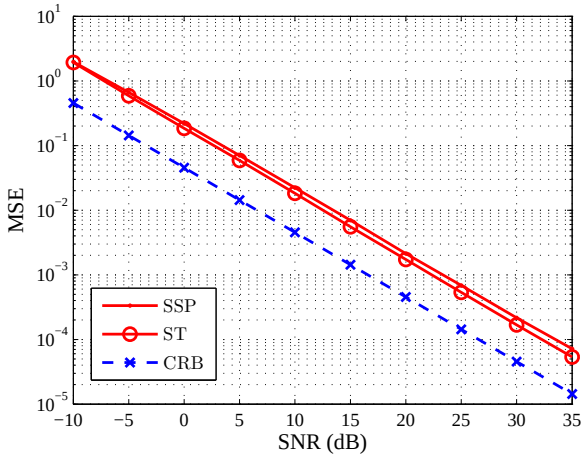


Fig. 6: The MSE and CRB of the channel estimation without soft information by signal subspace projection for the channel parameters' setting of *Case* 1.

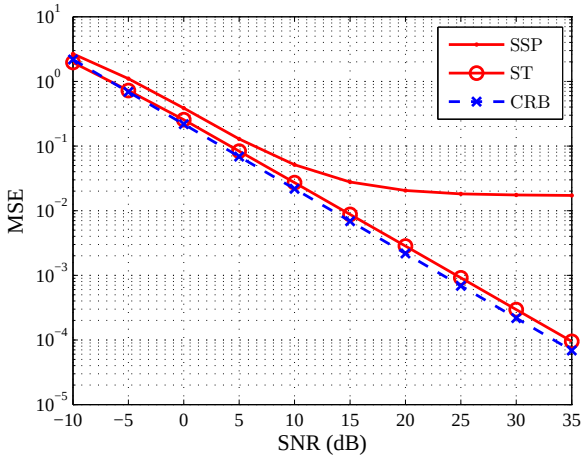
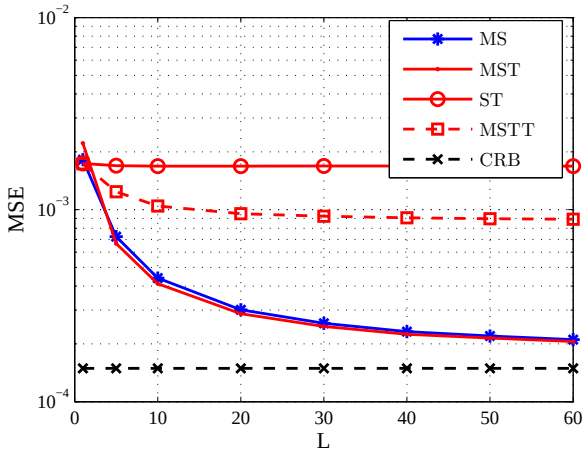
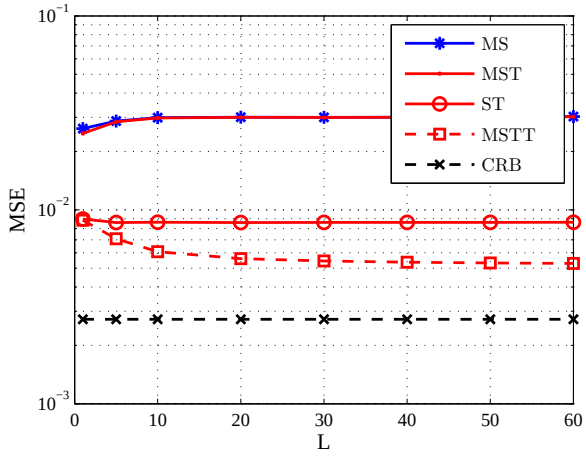


Fig. 7: The MSE and CRB of the channel estimation without soft information by signal subspace projection for the channel parameters' setting of *Case* 2.



(a) The channel parameters' setting of *Case* 1.



(b) The channel parameters' setting of *Case* 2.

Fig. 8: The MSE and CRB of the channel estimation without soft information for multiple slots

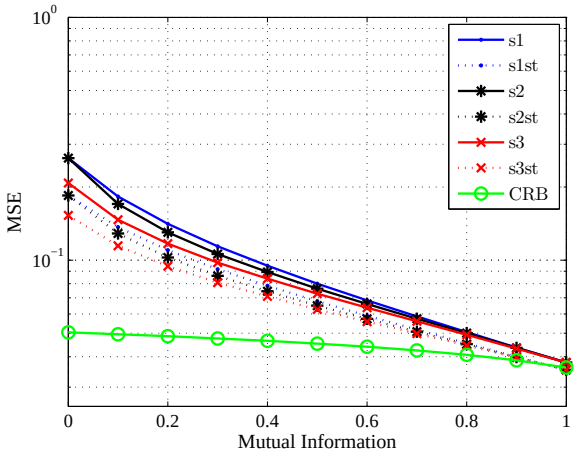


Fig. 9: The MSE of the channel estimates versus the mutual information for the channel parameters' setting of *Case* 2.

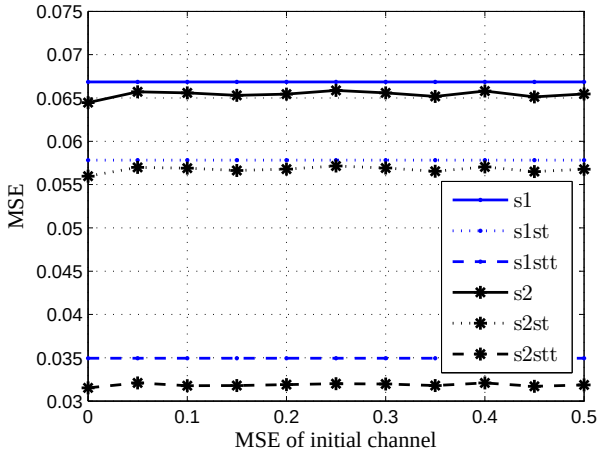


Fig. 10: Sensitivity of the soft based channel estimation methods "s1" and "s2" to the MSE of initial channel estimates for the channel parameters' setting of *Case* 2.

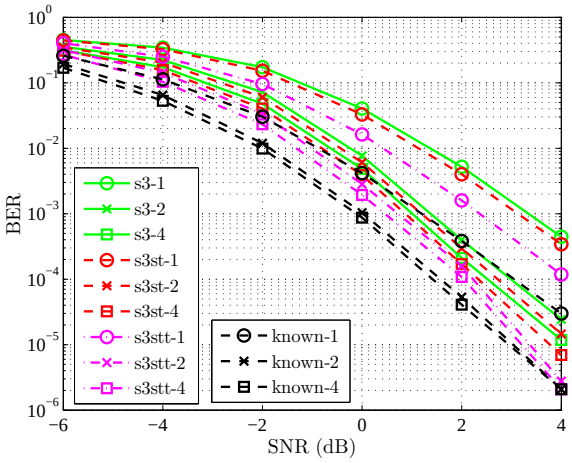


Fig. 11: The BER performance for the soft based channel estimates $\hat{\boldsymbol{H}}_{s3}$, $\hat{\boldsymbol{H}}_{s3st}$ and $\hat{\boldsymbol{H}}_{s3stt}$ versus SNR for varying number of iterations.

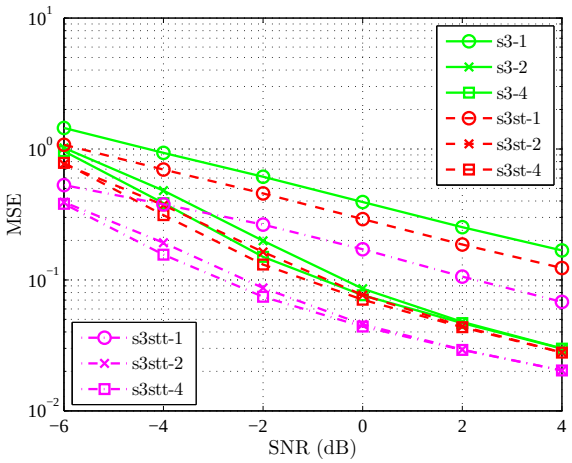


Fig. 12: The MSE of the soft based channel estimates $\hat{\boldsymbol{H}}_{s3}$, $\hat{\boldsymbol{H}}_{s3st}$ and $\hat{\boldsymbol{H}}_{s3stt}$ verse SNR for varying iterations.