electronics

MDPI

Article

Downlink Channel Estimation in Massive Multiple-Input Multiple-Output with Correlated Sparsity by Overcomplete Dictionary and Bayesian Inference

Wei Lu [1,*], Yongliang Wang [1], Xiaoqiao Wen [1], Shixin Peng [2] and Liang Zhong [3]

[1] Air Force Early Warning Academy, Wuhan 430019, China; ylwangkjld@163.com (Y.W.); evacassidy@126.com (X.W.)
[2] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China; pengshixin@mail.ccnu.edu.cn
[3] Department of communication system, China University of Geoscience, Wuhan 430074, China; hustzhongliang@gmail.com
* Correspondence: lvweiwhut@aliyun.com; Tel.: +86-027-85695800

Received: 24 March 2019; Accepted: 24 April 2019; Published: 28 April 2019

Abstract: We exploited the temporal correlation of channels in the angular domain for the downlink channel estimation in a massive multiple-input multiple-output (MIMO) system. Based on the slow time-varying channel supports in the angular domain, we combined the channel support information of the downlink angular channel in the previous timeslot into the channel estimation in the current timeslot. A downlink channel estimation method based on variational Bayesian inference (VBI) and overcomplete dictionary was proposed, in which the support prior information of the previous timeslot was merged into the VBI for the channel estimation in the current timeslot. Meanwhile the VBI was discussed for a complex value in our system model, and the structural sparsity was utilized in the Bayesian inference. The Bayesian Cramér–Rao bound for the channel estimation mean square error (MSE) was also given out. Compared with other algorithms, the proposed algorithm with overcomplete dictionary achieved a better performance in terms of channel estimation MSE in simulations.

Keywords: massive MIMO; channel estimation; Bayesian inference; overcomplete dictionary

1. Introduction

Massive multiple-input multiple-output (MIMO) is the key technology for next generation wireless communication. The large number of antennas enable high spectrum efficiency and lower power consumption [1]. To get these benefits, the base station (BS) needs to acquire the channel stated information (CSI) for uplink and downlink. Pilot-based channel estimation is widely used in wireless communication systems. In the time division duplex (TDD) system, the channel reciprocity is used to get the CSI by only estimating the uplink channel at BS. In the frequency division duplex (FDD) system, the channel reciprocity cannot be used directly. In FDD massive MIMO system it is challenging to get the downlink CSI with the conventional feedback scheme. In the conventional feedback scheme each user estimates its channel and then feeds back the estimated CSI to the BS. The pilot and feedback overheads are high for massive MIMO, since they are scaling linearly with the number of antennas. Hence, it is important to design an efficient downlink channel estimation and feedback scheme for a FDD massive MIMO system.

By exploiting the sparsity in massive MIMO channel, compressed sensing (CS) was applied in the channel estimation and feedback. The users could feed the compressed training measurements back to the BS, and an orthogonal matching pursuit (OMP) was used for downlink CSI recovery in [2]. In [3] the modified basis pursuit (MBP) was proposed by utilizing the partial priori signal support information to improve the recovery performance. In [4] the support information of a signal in the discrete fourier transform domain was incorporated into the weighted l_1 minimization approach for CS recovery, which could reduce the number of measurements by the size of the known part of support. In [5] a three-level weighting scheme based on the support information was used for the weighted l_1 minimization and the simulation results showed superiority. In [6] we exploited the reciprocity between uplink and downlink channels in the angular domain, and diagnosed the supports of the downlink channel from the estimated uplink channel, and proposed a weighted subspace pursuit (SP) channel estimation algorithm for FDD massive MIMO. It can be seen that CS was effective in the channel estimation for massive MIMO.

However, most of these algorithms need the sparsity level in the estimation algorithm, which is not practical in engineering scenarios. The Bayesian framework can be applied to the compressive channel estimation. In [7], Bayesian estimation of sparse massive MIMO channel was developed in which neighboring antennas shared among each other their information about the channel support. In [8] a variational expectation maximization strategy was used for massive MIMO channel estimation, and a Gaussian mixture prior model was designed to capture the individual sparsity for each channel and the joint sparsity among users. In [9] a sparse Bayesian learning algorithm was proposed for FDD massive MIMO channel estimation with arbitrary 2D-array. By the Bayesian framework in compressive channel estimation the sparsity level is unnecessary, and it has relatively better recovery performance. Additionally, there exists angular reciprocity in massive MIMO. For example, the channel covariance matrices for uplink and downlink are reconstructed by making use of the angle reciprocity between uplink and downlink channels in [10]. Hence it is promising to apply the angular reciprocity and Bayesian framework in the compressive massive MIMO channel estimation.

Additionally, there exists angular reciprocity in the FDD massive MIMO. There is also time correlation of channels. In [11] a differential compressive feedback in FDD massive MIMO was proposed based on the channel impulses response (CIR) between timeslots, which were slow time-varying and sparse, and the differential CIR between two CIRs in adjacent timeslots was sparse. Inspired by the sparsity in the angular domain and time correlation of channels, the correlated angular sparsity can also be exploited for massive MIMO channel estimation.

In this paper we proposed a downlink channel estimation in a TDD/FDD massive MIMO system. The timeslots were divided into groups. In each group the estimated channel support information of the previous timeslot was utilized by the following timeslot. The correlated angular sparsity between timeslots in the downlink channel was utilized in the Bayesian inference for channel recovery. We transformed the complex sparse vector to the real sparse vector recovery by Bayesian inference, and the structural sparsity of the transformed real sparse vector was utilized. Meanwhile, the prior support information from the estimated channel in the previous timeslot was made use of in modeling the hidden hyperparameters in the Bayesian model. A Bayesian Cramér–Rao bound analysis is presented, and simulations are given out to verify the performance of the proposed algorithm. The main contributions were as follows: (1) a group-based channel estimation scheme was proposed, in which previous estimated channel support information was used as the priori information in the following timeslot due to the sparsity correlation; (2) priori information was merged into the Bayesian inference algorithm for channel recovery; (3) the Bayesian Cramér–Rao bound for the channel estimation mean square error (MSE) was analyzed.

The system model is illustrated in Section 2, while the proposed channel estimation algorithm based on Bayesian inference is presented in Section 3. The Bayesian Cramér–Rao bound (BCRB) for the channel estimation of mean square error (MSE) is given out in Section 4. Simulations and conclusions are presented in Sections 5 and 6.

In the paper, we used the following notations. Scalars, vectors and matrices were denoted by lower-case, boldface lower-case and boldface upper-case symbols. The probability density function of a given random variable was denoted by $p(\cdot)$. Gamma$(x|a, b)$ was the Gamma probability density function (PDF) with shape parameters a and b for x, while Normal$(x|c, d)$ was the Gaussian PDF with parameters mean c and variance d for x. $\Gamma(\cdot)$ was the Gamma function, and $\ln(\cdot)$ was the logarithm function. $\mathrm{Tr}(\cdot)$ stood for the trace operator. $\mathbb{E}_{\mathrm{a}}(\cdot)$ denoted the expectation operation with the PDF of variable a.

2. System Model

We considered a massive MIMO TDD/FDD system with a single user, and assumed that the BS was equipped with N antennas and the user terminal (UT) had a single antenna. For the downlink channel estimation in the massive MIMO system, the BS transmitted the pilots to UT. The UT received the pilots and fed back the received signal to the BS directly. The received signal $\mathbf{y}^d(t)$ at the UT in the t-th timeslot was written as

$$\mathbf{y}^d(t) = \sqrt{\rho^d}\mathbf{A}\mathbf{h}^d(t) + \mathbf{n}^d(t) \tag{1}$$

where $\mathbf{h}^d(t) \in \mathbb{C}^{N\times 1}$ is the downlink channel, $\mathbf{A} \in \mathbb{C}^{T_d\times N}$ is the downlink pilots, T_d is the pilot length, ρ^d is the downlink received power, $\mathbf{n}^d \in \mathbb{C}^{T_d\times 1}$ is the received noise with each element to be i.i.d Gaussian with mean 0 and variance σ^2, $\mathbf{y}^d(t) \in \mathbb{C}^{T_d\times 1}$ is the received signal at UT.

Since in the massive MIMO there existed sparsity, when $\mathbf{D}^d \in \mathbb{C}^{N\times M}$ was the channel dictionary for downlink channel which could be unitary dictionary or overcomplete dictionary ($M > N$, their column vector had the form of steering vector with a different sampling angle), $\mathbf{h}_a^d(t)$ was the sparse representation with $\mathbf{h}^d(t) = \mathbf{D}^d\mathbf{h}_a^d(t)$. In this paper we applied the overcomplete dictionary to present the sparse angular channel to get a better recovery performance. In the downlink channel estimation, we needed to obtain $\hat{\mathbf{h}}_a^d(t)$ the estimated downlink channel in the angular domain in the t-th timeslot.

By utilizing the sparse channel representation we then had

$$\mathbf{y}^d(t) = \sqrt{\rho^d}\mathbf{A}\mathbf{D}^d\mathbf{h}_a^d(t) + \mathbf{n}^d(t) \tag{2}$$

For simplicity, the timeslot mark is omitted in the following equations. Since $\mathbf{y}^d(t)$, $\mathbf{h}_a^d(t)$, and $\mathbf{n}^d(t)$ are complex number vectors, we could rewrite Equation (2) into real number vectors as

$$\begin{bmatrix} \mathrm{Re}(\mathbf{y}^d) \\ \mathrm{Im}(\mathbf{y}^d) \end{bmatrix} = \begin{bmatrix} \mathrm{Re}(\sqrt{\rho^d}\mathbf{A}\mathbf{D}^d) & -\mathrm{Im}(\sqrt{\rho^d}\mathbf{A}\mathbf{D}^d) \\ \mathrm{Im}(\sqrt{\rho^d}\mathbf{A}\mathbf{D}^d) & \mathrm{Re}(\sqrt{\rho^d}\mathbf{A}\mathbf{D}^d) \end{bmatrix} \begin{bmatrix} \mathrm{Re}(\mathbf{h}_a^d(t)) \\ \mathrm{Im}(\mathbf{h}_a^d(t)) \end{bmatrix} + \begin{bmatrix} \mathrm{Re}(\mathbf{n}^d(t)) \\ \mathrm{Im}(\mathbf{n}^d(t)) \end{bmatrix} \tag{3}$$

where Re$(\cdot)$ and Im$(\cdot)$ denote the real and imaginary parts respectively. For simplicity, we rewrote Equation (3) as

$$\bar{\mathbf{y}} = \overline{\mathbf{A}}\bar{\mathbf{h}} + \bar{\mathbf{n}} \tag{4}$$

where $\bar{\mathbf{y}} = \begin{bmatrix} \mathrm{Re}(\mathbf{y}^d) \\ \mathrm{Im}(\mathbf{y}^d) \end{bmatrix}$, $\overline{\mathbf{A}} = \begin{bmatrix} \mathrm{Re}(\sqrt{\rho^d}\mathbf{A}\mathbf{D}^d) & -\mathrm{Im}(\sqrt{\rho^d}\mathbf{A}\mathbf{D}^d) \\ \mathrm{Im}(\sqrt{\rho^d}\mathbf{A}\mathbf{D}^d) & \mathrm{Re}(\sqrt{\rho^d}\mathbf{A}\mathbf{D}^d) \end{bmatrix}$, $\bar{\mathbf{h}} = \begin{bmatrix} \mathrm{Re}(\mathbf{h}_a^d(t)) \\ \mathrm{Im}(\mathbf{h}_a^d(t)) \end{bmatrix}$ and $\bar{\mathbf{n}} = \begin{bmatrix} \mathrm{Re}(\mathbf{n}^d(t)) \\ \mathrm{Im}(\mathbf{n}^d(t)) \end{bmatrix}$.

On the other hand, we considered the meaning of sparse angular channel representation $\mathbf{h}_a^d(t)$. If the transmission angles were allocated exactly at the sampling points in the channel dictionary $\boldsymbol{D}^d$, then the corresponding coefficient in the $\mathbf{h}_a^d(t)$ was nonzero. If the path number was smaller than the antenna number, then $\mathbf{h}_a^d(t)$ was sparse. However, there was leakage effect induced by dictionary mismatch which will have deteriorated the sparsity of the angular channel representation [12]. When the movement velocity of UT was not very high, e.g., v = 12 km/h, and the typical timeslot duration τ = 0.5 ms, the movement distance of UT in one timeslot was 0.017 m. When the distance of UT and BS was 200 m, the angle change for the line of sight (LoS) transmission in one timeslot was 0.0049° which was much smaller than the sampling interval in the dictionary. For the non-LoS (NLoS) transmission, the angle change was also small which is discussed in Section 4.1. Hence the transmission angle change between two timeslots is very small if the transmission environment doesn't change dramatically, and there is correlation in the angular channel sparsity between adjacent timeslots. In other words, the information regarding the estimated angular channel in the previous timeslot could be utilized in the current channel estimation.

It was proven that the prior support information could improve the channel recovery performance [3–6]. Hence in this paper we made use of the prior support information from the previous timeslot to improve the Bayesian channel estimation. In the following section we have discussed how to merge the prior information into the Bayesian inference algorithm for channel estimation.

3. Proposed Algorithm

We designed a three-layer hierarchical graphical model as shown in Figure 1. In the first layer, $\overline{\mathbf{h}}$ was assigned a Gaussian prior distribution

$$p(\overline{\mathbf{h}}|\boldsymbol{\alpha}) = \prod_{i=1}^{2N} p(\overline{h}_i \big| \alpha_i) \tag{5}$$

where $\overline{h}_i$ and α_i are the i-th entry in $\overline{\mathbf{h}}$ and $\boldsymbol{\alpha}$ respectively, $p(\overline{h}_i \big| \alpha_i) = Normal(\overline{h}_i \big| 0, \alpha_i)$ and α_i is the inverse variance of the Gaussian distribution. When $\overline{h}_i$ is close to 0, then α_i is very large, and vice versa.

In the second layer, we assumed a Gamma distribution as hyperpriors over the hyperparameters α_i, and it can be presented as

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{2N} Gamma(\alpha_i \big| a_i, b_i) \tag{6}$$

where Gamma(·) is the Gamma PDF, and the parameters a_i and b_i characterize the shape of Gamma PDF. For fixed a_i, the larger b_i is, the smaller α_i is; then $\overline{h}_i$ tends to be nonzero. In the sparse Bayesian learning a_i and b_i were set to be very small for non-informative hyperprior over α_i [13].

In our model, we set a_i to be constant with a predefined value, and we modeled b_i as random parameters. In Figure 1 it could be found that the entries of $\overline{\boldsymbol{y}}$ were divided into two sets by their indices, i.e., $\{S, S_{+N}\}$ and $\{S, S_{+N}\}^c$, where S was the set with channel support indices from the previous timeslot, and S_{+N} was the set with each index in S added by N, since we converted the complex system model to the real system model as Equation (3).$\{S, S_{+N}\}^c$ was the complementary set of $\{S, S_{+N}\}$. For example, in the $(t-1)$-th timeslot, the positions of nonzero entries or called supports in $\mathbf{h}_a^d(t-1)$ were S = {4, 5, 6}, then $S_{+N} = \{4+N, 5+N, 6+N\}$. The probable supports for $\mathbf{h}_a^d(t)$ in the current t-th timeslot can be assumed to be the same as those for previous $(t-1)$-th timeslot for simplicity. On the other hand, we could have also diagnosed the probable channel supports further by taking the angle deviation and leakage effects into consideration. In this paper we adopted the support diagnosis algorithm, and the details can be found in [6].

For $\overline{y}_j$, $j \in \{S, S_{+N}\}$, we employed a Gamma distribution over the hyperparameters b_i in the third layer as

$$\text{Gamma}(b_i|c, d) = \Gamma(c)^{-1} d^c b_i^{c-1} e^{-db_i} \tag{7}$$

where c and d characterize the shape of Gamma PDF. By the system model and assumptions for massive MIMO, we could use a Bayesian inference to perform the sparse channel recovery.

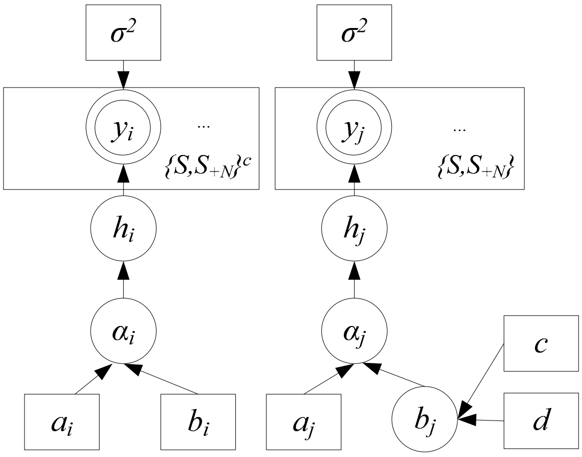

Figure 1. Graphical model for the channel estimation with Bayesian inference. The nodes with double circle, single circle and square correspond to the observed data, hidden variables and parameters, respectively.

According to the standard Bayesian inference [14], let $\mathbf{z} \triangleq \left\{\overline{\boldsymbol{h}}, \boldsymbol{\alpha}, \boldsymbol{b}\right\}$, we have

$$\begin{aligned}\ln p(\mathbf{z}_i) &= \mathbb{E}_{\mathbf{z}_i, i\neq j}[\ln p(\overline{\mathbf{y}},\mathbf{z})] + \text{constant}\\ &\propto \mathbb{E}_{\mathbf{z}_i, i\neq j}[\ln p(\overline{\mathbf{y}},\mathbf{z})]\end{aligned} \tag{8}$$

where constant is a constant used for $p(\mathbf{z}_i)$ normalization, $p(\overline{y}, z)$ is the joint pdf for $\overline{\boldsymbol{h}}$ and $\mathbf{z}$, and $\boldsymbol{z_i}$ can be $\overline{\boldsymbol{h}}, \boldsymbol{\alpha},$ and$\boldsymbol{b}$. We have $p(\overline{\mathbf{y}}, \mathbf{z}) = p(\mathbf{z}|\overline{\mathbf{y}})p(\overline{y})$. We assume $p(\mathbf{z}|\overline{\mathbf{y}})$ posterior independence among the hidden variables **z**, then $p(\mathbf{z}|\overline{\mathbf{y}}) \approx p(\mathbf{z})$, and $p(\mathbf{z})$ is the product of PDF of $\overline{\boldsymbol{h}}, \boldsymbol{\alpha},$ and $\boldsymbol{b}$.

In order to make use of the prior support information from the previous timeslot and the structure sparsity in Equation (4), we needed to make some modifications to the standard Bayesian inference. The main considerations for the modifications were as follows:

(I) Since we rewrote Equation (2) as Equation (4), if $h^d_{a,i}$ was nonzero, then $\overline{h}_i$ and $\overline{h}_{i+N}$ were nonzero simultaneously. Hence it was wise to assume that b_i and b_{i+N} were the same;

(II) In the standard Bayesian learning a_i and b_i were set to be very small for non-informative hyperprior over α_i. This assumption was valid if no prior information was provided. If the prior support information was available, such as that the support information of the previous timeslot could be used for channel estimation in the coming timeslot by sparsity correlation, it was wise to assume that the supports between adjacent timeslots were partially common. If the *i-th* element in the angular channel vector was nonzero, then the hyperparameter b_i and b_{i+N} tended to be variables rather than to be fixed small numbers, which meant only for the indices from the prior support set S the third layer prior model was adopted.

It can be seen that the consideration (II) was similar to [15]. However, our proposed algorithm was extended for a complex number system and the structure sparsity was considered. However, on the other hand, the overcomplete dictionary was adopted in our algorithm.

The proposed uplink-aided downlink channel estimation based on Bayesian inference was as follows:

(i) Update of p($\overline{h}$)

According to Equation (8), by ignoring the terms which are independent of $\overline{\boldsymbol{h}}$, we have

$$\begin{aligned}\ln p(\overline{\mathbf{h}}) &\propto \mathbb{E}_{\boldsymbol{\alpha},\mathbf{b}}\Big[\ln p(\overline{\mathbf{y}}\big|\overline{\mathbf{h}}) + \ln p(\overline{\mathbf{h}}\big|\boldsymbol{\alpha}) + \ln p(\mathbf{b})\Big] \\ &\propto \mathbb{E}_{\boldsymbol{\alpha},\mathbf{b}}\Big[\ln p(\overline{\mathbf{y}}\big|\overline{\mathbf{h}}) + \ln p(\overline{\mathbf{h}}\big|\boldsymbol{\alpha})\Big] \\ &= \tfrac{-1}{2\sigma^2}(\overline{\mathbf{y}} - \overline{\mathbf{A}\mathbf{h}})^T(\overline{\mathbf{y}} - \overline{\mathbf{A}\mathbf{h}}) - \tfrac{1}{2}\overline{\mathbf{h}}^T\boldsymbol{\Lambda}\overline{\mathbf{h}}\end{aligned} \tag{9}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}\{\mathbb{E}_{\alpha}[\alpha_i]\}$, σ^2 is the noise variance in the system model, the vectors $\mathbf{b}$ and $\boldsymbol{\alpha}$ are comprised by b_i and α_i respectively. Since $p(\overline{\mathbf{y}}|\overline{\mathbf{h}})$ and $p\left(\overline{h}\middle|\boldsymbol{\alpha}\right)$ are a Gaussian distribution, then $p\left(\overline{h}\right)$ follows a Gaussian distribution with the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\phi}$ given by

$$\upmu = \frac{1}{\upsigma^2}\boldsymbol{\Phi}\overline{\mathbf{A}}^T\overline{\mathbf{y}} \tag{10}$$

$$\boldsymbol{\Phi} = \left(\frac{1}{\upsigma^2}\overline{\mathbf{A}}^T\overline{\mathbf{A}} + \boldsymbol{\Lambda}\right) \tag{11}$$

(ii) Update of p(α)

According to Equation (8), by ignoring the terms which are independent of $\boldsymbol{\alpha}$, we have

$$\begin{aligned}\ln p(\boldsymbol{\alpha}) &\propto \mathbb{E}_{\overline{\mathbf{h}},\mathbf{b}}\Big[\ln p(\overline{\mathbf{y}}\big|\overline{\mathbf{h}}) + \ln p(\overline{\mathbf{h}}\big|\boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha}|\mathbf{a},\mathbf{b}\,) + \ln p(\mathbf{b})\Big] \\ &\propto \mathbb{E}_{\overline{\mathbf{h}},\mathbf{b}}\Big[\ln p(\overline{\mathbf{h}}\big|\boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha}|\mathbf{a},\mathbf{b}\,)\Big] \\ &= \sum_{i=1}^{2N}\mathbb{E}_{\overline{h},b}\Big\{(a_i - 0.5)\ln\alpha_i - \Big(0.5\overline{h}_i^2 + b_i\Big)\alpha_i\Big\} \\ &= \sum_{i\in\{S,S_{+N}\}}\mathbb{E}_{\overline{h},b}\Big\{(a_i + 0.5 - 1)\ln\alpha_i - \Big(0.5\overline{h}_i^2 + b_i\Big)\alpha_i\Big\} + \\ &\sum_{i\in\{S,S_{+N}\}^c}\mathbb{E}_{\overline{h},b}\Big\{(a - 0.5)\ln\alpha_i - \Big(0.5\overline{h}_i^2 + b_i\Big)\alpha_i\Big\} \\ &= \sum_{i\in\{S,S_{+N}\}}\left\{(a_i + 0.5 - 1)\ln\alpha_i - \left(\frac{\mathbb{E}_{\overline{h},b}(b_i+b_{i+N})}{2} + \frac{\mathbb{E}_{\overline{h},b}(\overline{h}_i^2+\overline{h}_{i+N}^2)}{4}\right)\alpha_i\right\} + \\ &\sum_{i\in\{S,S_{+N}\}^c}\left\{(a_i + 0.5 - 1)\ln\alpha_i - \left(b_i + \frac{\mathbb{E}_{\overline{h},b}(\overline{h}_i^2+\overline{h}_{i+N}^2)}{4}\right)\alpha_i\right\}\end{aligned} \tag{12}$$

where S is the estimated support set from the previous timeslot. Since the complex system model was converted in Equation (4). By (II), $\boldsymbol{S}_{+N} \triangleq \{s_i + N\}$ was also the support set in the converted system model in Equation (4). For $i \in \{S, S_{+N}\}$, b_i is variable number, b_i and b_{i+N} were assumed to be the same, we used $0.5\mathbb{E}_{\overline{\boldsymbol{h}},\boldsymbol{b}}(b_i + b_{i+\mathrm{N}})$ to present $\mathbb{E}_{\overline{\boldsymbol{h}},\boldsymbol{b}}(b_i)$. The same assumption was applied to $\overline{h}_i$ and $\overline{h}_{i+N}$ with $\mathbb{E}_{\overline{\mathbf{h}},\mathbf{b}}(\overline{h}_i^2) = 0.5\mathbb{E}_{\overline{\mathbf{h}},\mathbf{b}}(\overline{h}_i^2 + \overline{h}_{i+N}^2)$. In this way the structural sparsity was utilized.

Since $p(\boldsymbol{\alpha}|\boldsymbol{a},\boldsymbol{b}\,)$ is the Gamma distribution and $p\left(\overline{\boldsymbol{h}}\middle|\boldsymbol{\alpha}\right)$ is the Gaussian distribution, $p(\boldsymbol{\alpha})$ is the Gamma distribution. Then $p(\alpha_i)$ is also the Gamma distribution with the updated parameters $\widetilde{a}_i$ and $\widetilde{b}_i$ given by

$$\widetilde{a}_i = a_i + 0.5 \tag{13}$$

$$\widetilde{b}_i = \begin{cases} \frac{\mathbb{E}_{\overline{\mathbf{h}},\mathbf{b}}(b_i+b_{i+N})}{2} + \frac{\mathbb{E}_{\overline{\mathbf{h}},\mathbf{b}}(\overline{h}_i^2+\overline{h}_{i+N}^2)}{4}, i \in \{S, S_{+N}\} \\ b_i + \frac{\mathbb{E}_{\overline{\mathbf{h}},\mathbf{b}}(\overline{h}_i^2+\overline{h}_{i+N}^2)}{4}, i \in \{S, S_{+N}\}^c \end{cases} \tag{14}$$

(iii) Update of $p(\mathbf{b}_{\{S,S_{+N}\}})$

According to Equation (8), by ignoring the terms which are independent of $\mathbf{b}$, we have

$$\begin{aligned} \ln p(\mathbf{b}_{\{S,S_{+N}\}}) & \propto \mathbb{E}_{\boldsymbol{\alpha},\overline{\mathbf{h}}}\Big[\ln p(\overline{\mathbf{y}}\Big|\overline{\mathbf{h}}) + \ln p(\overline{\mathbf{h}}\Big|\boldsymbol{\alpha}) + \ln p(\boldsymbol{\alpha}\Big|\mathbf{a},\mathbf{b}) + \ln p(\mathbf{b}\Big|\mathbf{c},\mathbf{d})\Big] \\ & \propto \mathbb{E}_{\boldsymbol{\alpha},\overline{\mathbf{h}}}[\ln p(\boldsymbol{\alpha}\Big|\mathbf{a},\mathbf{b}) + \ln p(\mathbf{b}\Big|\mathbf{c},\mathbf{d})] \\ & = \sum_{i\in\{S,S_{+N}\}} \{-b_i\mathbb{E}_{\boldsymbol{\alpha}}(\alpha_i) + (c_i-1)\ln b_i - d_i b_i\} \end{aligned} \tag{15}$$

where $\mathbf{b}_{\{S,S_{+N}\}}$ is comprised by the entries indicated by $\{S,S_{+N}\}$ in $\mathbf{b}$. In (15) the α, $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{d}$ are also comprised by their indicated $\{S,S_{+N}\}$, the subscript $\{S,S_{+N}\}$ is omitted for simplicity. As shown in Figure 1, $\mathbf{b}_{\{S,S_{+N}\}}$ was modelled as a Gamma distribution. Since $p(\alpha_i|a_i,b_i)$ and $p(b_i|c_i,d_i)$ were a Gamma distribution, $p(\tilde{\mathbf{b}}_{\{S,S_{+N}\}})$ was Gamma$(\widetilde{b}_{i\in\{S,S_{+M}\}}\Big|\widetilde{c_i},\widetilde{d_i})$, and the updated $\widetilde{c_i}$ and $\widetilde{d_i}$ were given by

$$\widetilde{c_i} = a_i + c_i \tag{16}$$

$$\widetilde{d_i} = d_i + \mathbb{E}_{\alpha}(\alpha_{\mathrm{i}}) \tag{17}$$

Then the Bayesian inference for the channel estimation was executed iteratively among (i), (ii), and (iii). The details of the algorithm are summarized in step 3 of Algorithm 1. When the estimated channel vector $\overline{\mathbf{h}\prime}$ was recovered, we needed to convert it to the complex vector $\boldsymbol{h}_a^d$ according to Equation (3).

Algorithm 1 Downlink channel estimation with variational inference algorithm and overcomplete dictionary.

```
Input: $\overline{\mathbf{A}}, \overline{\mathbf{y}}, \sigma^2$
Output: $\overline{\mathbf{h}\prime}$
1.  Divide the timeslots into groups, and with each group comprised by $t_g$ timeslots.
2.  For the first timeslot in the group, use variational Bayesian inference (VBI) for channel estimation, and obtain the angular channel supports.
3.  For the rest of the timeslots in the group, utilize the support information from the previous timeslot for channel estimation one by one. The recovery algorithm in each timeslot is as follows:
    3.1.  Initialize $\boldsymbol{\alpha}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$.
    3.2.  $\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\phi}\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{y}}, \boldsymbol{\phi} = (\frac{1}{\sigma^2}\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}} + \boldsymbol{\Lambda}), \mathbb{E}_{\overline{\mathbf{h}},\mathbf{b}}(\overline{\mathbf{h}}_{\mathrm{i}}^2) = \mu_{\mathrm{i}}^2 + \phi_{\mathrm{i,i}}$, where $\boldsymbol{\Lambda} = \mathrm{diag}\{\mathbb{E}_{\boldsymbol{\alpha}}[\alpha_i]\}$, $\mu_i$ is the i-th entry in $\boldsymbol{\mu}$, and $\phi_{\mathrm{i,i}}$ is the i-th diagonal entry in $\boldsymbol{\phi}$.
    3.3.  Update $\widetilde{a_i}$ and $\widetilde{b_i}$ according to Equations (13) and (14) in (ii) ($\widetilde{a_i}$ and $\widetilde{b_i}$ are the updated $a_i$ and $b_i$, and $a_i$ and $b_i$ are the results from last iteration); then according to the property of the Gamma distribution variable, $\mathbb{E}_{\boldsymbol{\alpha}}(\alpha_i) = \widetilde{a_i}/\widetilde{b_i}$.
    3.4.  Update $\widetilde{c}$ and $\widetilde{d}$ according to Equations (16) and (17) in (iii) ($\widetilde{c}$ and $\widetilde{d}$ are the updated $c$ and $d$, and $c$ and $d$ are the results from last iteration); then according to the property of the Gamma distribution variable, $\mathbb{E}_{\boldsymbol{\alpha}}(\widetilde{b_i}) = \widetilde{c}/\widetilde{d}$.
    3.5.  Go to step 3.2 until stop criteria meets.
    3.6.  Then $\overline{\boldsymbol{h}\prime} = \boldsymbol{\mu}$.
4.  Go back to step 2 for a new group of timeslots.
```

In a practical massive MIMO system, the transmission environment may change suddenly, in this way the correlation of sparsity between adjacent timeslots will deteriorate, and the previous channel support information cannot be utilized. On the other hand, the error will accumulate if the previous channel support information is utilized timeslot by timeslot. Hence, the initialization is important for the robustness and efficiency of the algorithm. As shown in Figure 2 divided the timeslots into groups, and each group was comprised of several timeslots. During the channel estimation for each group, the VBI was used for the channel estimation in the first timeslot, and then the proposed algorithm was executed for the remaining timeslots in which the channel support information of the previous timeslot was made use of by the current timeslot. This procedure is detailed in steps 1, 2 and 4 in Algorithm 1.

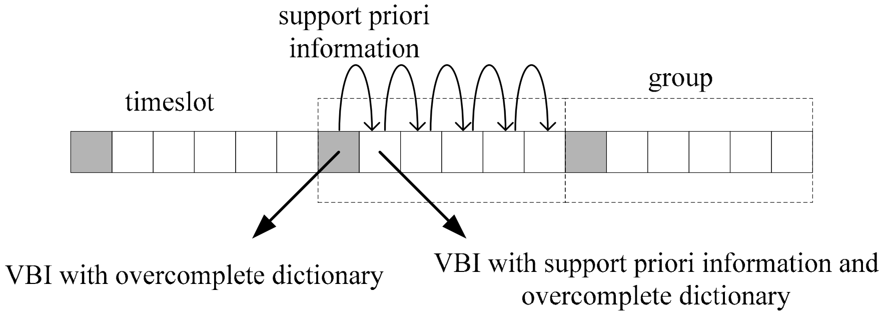
support priori information
timeslot
group
VBI with overcomplete dictionary
VBI with support priori information and overcomplete dictionary

Figure 2. Channel estimations by group. Each block represents one timeslot, and the block filled with grey is the timeslot with variational Bayesian inference (VBI) for the channel estimation, while the blank blocks are the timeslots with the proposed algorithm for channel estimation.

4. Discussion

4.1. Sparsity Correlation Analysis

The UT movement distance was very small when the velocity of UT was small and the timeslot was 0.5 ms. The reflector for the transmission was static during the UT moving between timeslots. The ellipse geometry channel model is shown in Figure 3. The line of sight (LoS) distance between BS and UT was d_{Los}, the non-LoS (NLoS) distance by reflector between BS and UT was d_{NLoS}, and the UT movement distance in one timeslot was d_Δ. If the transmission path was still reflected by the same reflector as shown in Figure 3, the maximum and minimum NLoS distances from BS to UT between timeslots were $d_{NLoS} + d_\Delta$ and $d_{NLoS} - d_\Delta$. The transmission angle change was Δ_θ. The distance between the reflector and BS was d_1. By some mathematical manipulations shown in Appendix A, we got

$$\Delta_\theta \approx \frac{2d_\Delta(d_{NLoS} - d_1)}{2d_1 d_{LoS}} \frac{1}{\sqrt{1 - \cos^2\theta}} \tag{18}$$

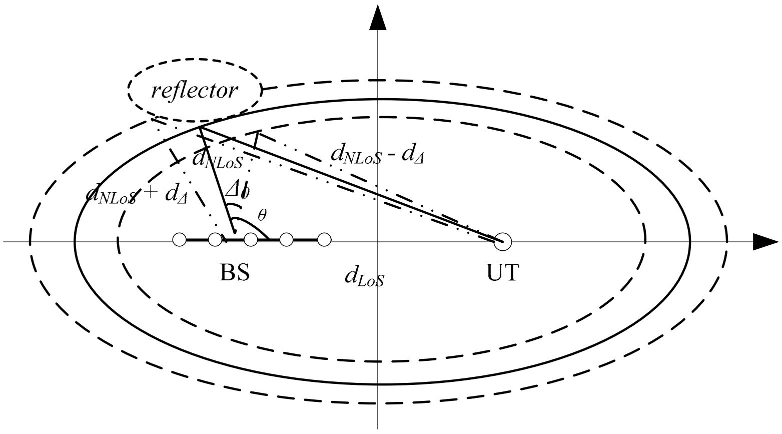

Figure 3. Ellipse geometry channel model for line of sight (LoS) and non-LoS (NLoS) transmission.

In order to illustrate the angle change Δ_θ during one timeslot, we assumed that $\mathrm{d}_{\mathrm{NLoS}}$ was 800 m, the velocity of UT was 14.4 km/h, and the typical timeslot duration τ = 0.5 ms, then the movement distance of UT in one timeslot was 0.02 m. By changing the distance between BS and reflector, as shown in Figure 4, the angle change was not more than 0.025°. It should be noted that when the LoS distance and the d_{NLoS} were fixed, BS and reflector distance could not be arbitrary vales due to triangle inequality. Hence, the angle of arrival or departure changed slowly and then there was sparsity correlation among the angular channels for adjacent timeslots.

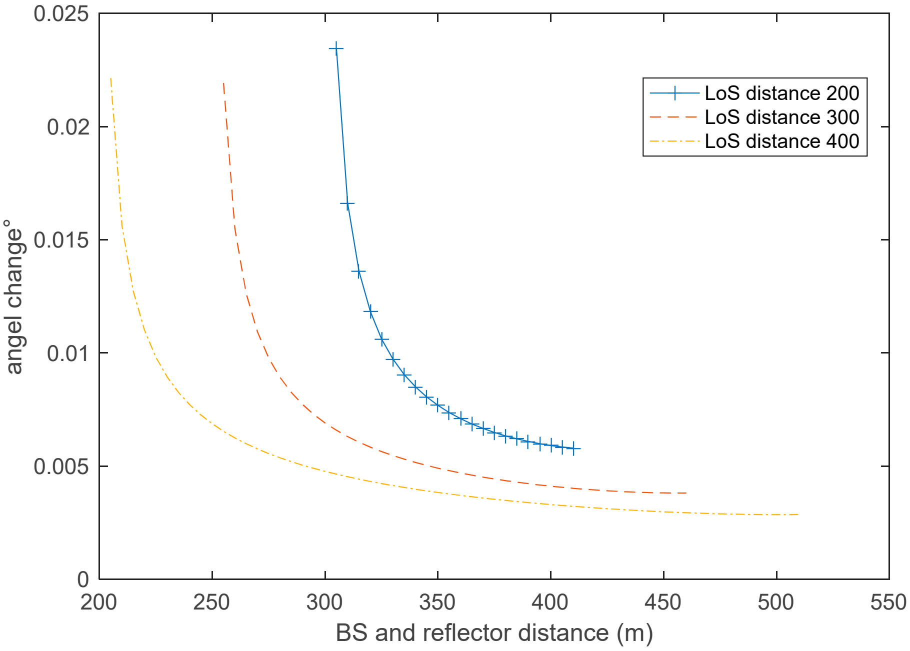

Figure 4. Transmission angle change during one timeslot with a different LoS distance and different distances between the base station (BS) and reflector.

4.2. Bayesian Cramér-Rao Bound Analysis

In this section we have discussed the Bayesian Cramér–Rao bound (BCRB) for the channel estimation with the proposed algorithm. Let $\mathbf{z} \triangleq \left\{\overline{h}, \sigma\right\}$, then the BCRB for the channel vector $\overline{\mathbf{h}}$ is given by the inverse of the Fisher information matrix $\mathbf{J}$ as:

$$\mathbf{J} = \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\overline{\mathbf{y}}, \mathbf{z})}{\partial z_i \partial z_j}\right\} \tag{19}$$

According to the system model in Section 2, $\overline{h}, \sigma$ are independent, the Fisher information matrix $\mathbf{J}$ is block diagonal. We can rewrite $p(\overline{y}, z)$ as

$$p(\overline{\mathbf{y}}, \mathbf{z}) = p(\overline{\mathbf{y}}|\mathbf{z})p(\overline{\mathbf{h}}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}|\mathbf{b})p(\mathbf{b}) \tag{20}$$

Then the BCRB on the MSE of the estimated channel vector $\overline{\mathbf{h}\prime}$ is given by

$$\mathbb{E}\left\{||\overline{\mathbf{h}}^{'} - \overline{\mathbf{h}}||^2\right\} \geq tr\left(\mathbf{J}^{-1}_{\overline{h}_i\overline{h}_j}\right) \tag{21}$$

where $\mathbf{J}_{\overline{h}_i\overline{h}_j} = \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\overline{\mathbf{y}}, \mathbf{z})}{\partial \overline{h}_i \partial \overline{h}_j}\right\}$ is the fisher information sub-matrix. Thus, we can obtain the Bayesian Cramér–Rao bound of the minimum mean square error for the estimated channel $\overline{\mathbf{h}\prime}$ as shown in Proposition 1.

Proposition 1. *The BCRB of MSE for the channel estimation $\overline{\boldsymbol{h}\prime}$ is represented as*

$$\mathbb{E}\left\{\|\overline{\mathbf{h}'}-\overline{\mathbf{h}}\|^2\right\} \geq tr\left(\left(diag(\mathbb{E}\left(\frac{1}{\alpha_i}\right)) + \frac{1}{\sigma^2}\overline{\mathbf{A}}^T\overline{\mathbf{A}}\right)^{-1}\right) = \sum_{i\in S}\frac{1}{\frac{1+c}{ad_i}+\frac{\lambda_i}{\sigma}} + \sum_{i\notin S}\frac{1}{\frac{b_i}{a}+\frac{\lambda_i}{\sigma}} \tag{22}$$

where S is the diagnosed support set, λ_i is the eigenvalues of $\overline{\boldsymbol{A}}^T\overline{\boldsymbol{A}}$, and $\overline{\boldsymbol{A}}^T\overline{\boldsymbol{A}} \in \mathbb{R}^{2N\times 2N}$, and a, b_i, c, and d_i are the parameters in the Bayesian model in Figure 1. *When $T_d, M \rightarrow \infty$ and $\frac{T_d}{M} = \beta$,* according to the random matrix theory, we have

$$\begin{aligned}\mathbb{E}\left\{\left(\overline{\mathbf{h}}-\hat{\overline{\mathbf{h}}}\right)^H\left(\overline{\mathbf{h}}-\hat{\overline{\mathbf{h}}}\right)\right\} \quad &\geq |S|\cdot\frac{1}{|S|}\sum_{i\in S}\frac{1}{\frac{1+c}{a\min(d)}++\frac{\lambda_i}{\sigma}} + (N-|S|)\cdot\frac{1}{(N-|S|)}\sum_{i\notin S}\frac{1}{\frac{\max(b)}{a}+\frac{\lambda_i}{\sigma}} \\ &\rightarrow |S|\frac{a\min(\mathbf{d})}{1+c}\left(1-\frac{F(snr_1,\beta)}{4\beta snr_1}\right) + (N-|S|)\frac{a}{\max(\mathbf{b})}\left(1-\frac{F(snr_2,\beta)}{4\beta snr_2}\right)\end{aligned} \tag{23}$$

*where $snr_1 = \frac{amin(\boldsymbol{d})}{(1+c)\sigma}$, $snr_2 = \frac{a}{\sigma max(\boldsymbol{b})}$, $F(x,z) = \left(\sqrt{x\left(1+\sqrt{z}\right)^2+1} - \sqrt{x\left(1-\sqrt{z}\right)^2+1}\right)^2$, $min(\boldsymbol{d})$ and $max(\boldsymbol{b})$ are the minimum and maximum entries in **d** and **b**.*

The proof of proposition 1 is presented in Appendix B. From proposition 1, we can see that the MSE lower bound is related to the priori support size $|S|$, $(1+c)/\min(\mathbf{d})$ and $\max(\mathbf{b})$ for the massive MIMO channel estimation.

5. Simulations

In the simulation, the support diagnosis algorithm in [6] was adopted, and we assumed that the transmission angle change between timeslots was within 1 degree. The pilot length was 50, and antenna number at the BS was 100. The channel was generated according to the spatial model as defined in 3GPP TR25.996. We compared our proposed algorithm with a unitary dictionary with a size of 100 and the overcomplete dictionary with a size of 150, 200, and 250, and compared this with a Bayesian sparse learning (SL) [16], weighted subspace pursuit (WSP) [6], weighted l_1 minimization (W-l_1 min) [5], weighted iteratively reweighted least square(W-IRLS), IRLS [17], compressive sampling matched pursuit (COSAMP) in [11], and l_1 minimization (l_1 min) [18].

In order to evaluate the channel estimation performance, we used a normalized mean-square error (MSE) between true and estimated channel vectors as follows:

$$MSE = \frac{1}{T}\sum_{T}\frac{\left\|\hat{h}^d - h^d\right\|^2}{\left\|h^d\right\|^2} \tag{24}$$

where T is the number of trials, $\hat{\mathbf{h}}^d$ and $\mathbf{h}^d$ are the estimated and original channel vector, respectively for each trial. In the simulations the trial number T was 250.

In Figure 5 the overcomplete dictionary size was 150 in the proposed algorithm. It could be seen that when the unitary dictionary was used, our proposed algorithm outperformed WSP, COSAMP and IRLS, but was a little worse than W-l_1 with a small gap. However, when the overcomplete dictionary was used, our proposed algorithm outperformed other algorithms, but almost had the same performance as SL with a little performance improvement which could be seen in the zoomed-in subfigure. The overcomplete dictionary in the proposed algorithm can dramatically improve the MSE performance due to the fact that there are more atoms in the overcomplete dictionary than in the unitary dictionary which can improve the sparsity in the angular channel; however, it doesn't mean that the larger the overcomplete dictionary size is, the better performance it has, which is shown in Figure 6.

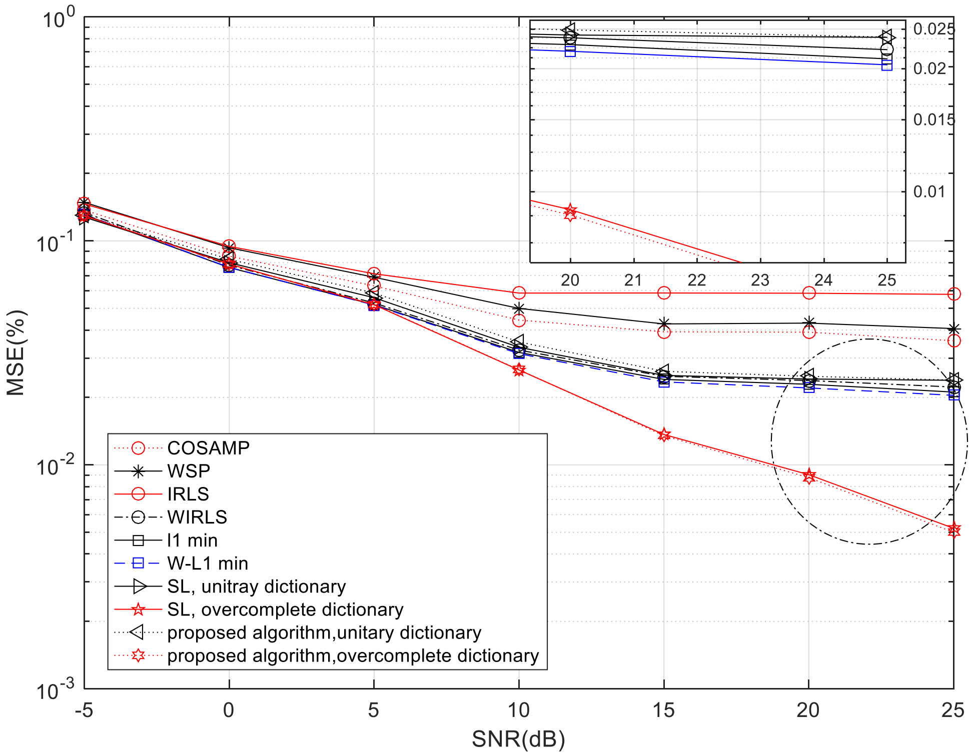

Figure 5. Comparisons of channel estimation mean square error (MSE) for different algorithms.

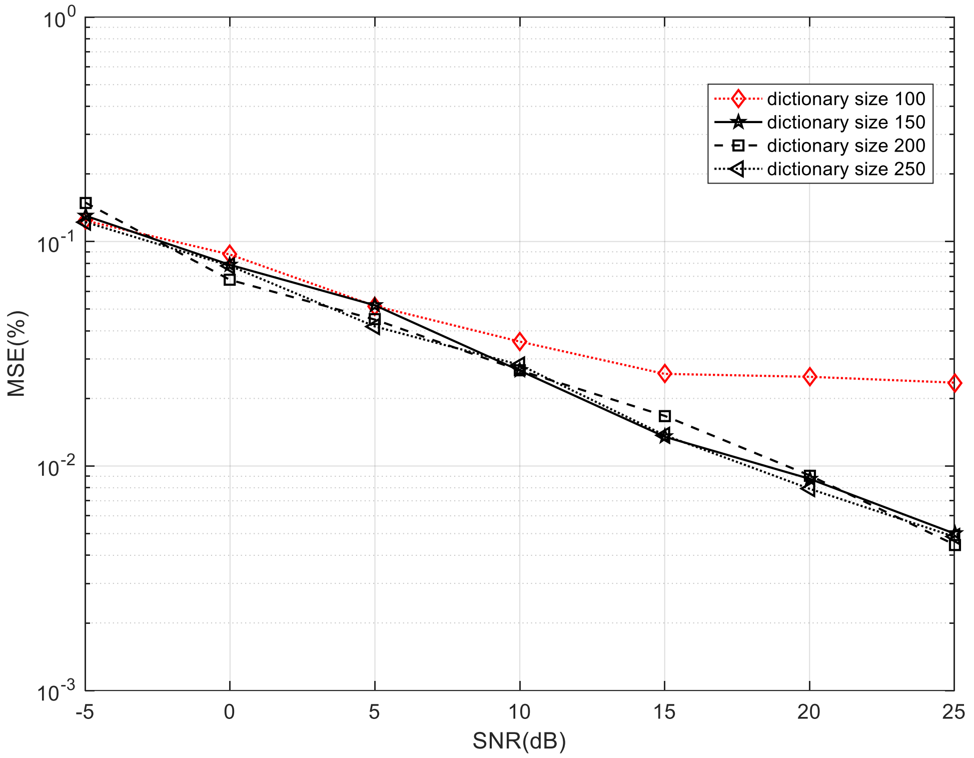

Figure 6. Comparisons of channel estimation of MSE for the proposed algorithm with different dictionary sizes.

We compared the performance of the proposed algorithm with different dictionary sizes in Figure 6. It can be seen that in the high SNR region the performance improved when an overcomplete dictionary was used, but the MSE performance gain did not improve when increasing the dictionary size. For example, the algorithm with a dictionary size of 150 had a relatively better performance than with a dictionary size of 100. However, the performances with a dictionary size of 200 and 250 almost gave the same trends as that with a dictionary size of 150. This was because the larger dictionary would induce angel ambiguity because the correlation of atoms increased. Hence, in the practical engineering, the dictionary size is not recommended to be very large. A large dictionary size is computationally expensive and the benefit is limited. It also should be noted that in the low SNR region the MSE performance with a larger dictionary size did not always do better than those with a smaller dictionary size. For example, when the SNR was 0 dB, they hadsimilar performance. The reason was that in the low SNR region the estimated channel support of the previous timeslot was not accurate enough, and on the other hand larger dictionary size would have deteriorated the dictionary incoherence.

We compared the runtime and convergence performance of the proposed algorithm with a different dictionary size in Figure 7. The relative error was defined as the ratio of the difference of adjacent iteration results to the previous iteration result. It can be seen that the proposed algorithm with dictionary size 150 converged fast than with a dictionary size of 100. However, the improvement had its price, and the runtime for the proposed algorithm with dictionary size 150 was longer which meant that the computational complexity was higher with a larger dictionary size. Based on the simulation results shown in Figures 6 and 7, when the antenna at BS is 100, the dictionary size is recommended to be set at 150 or so to balance the performance improvement and computation complexity.

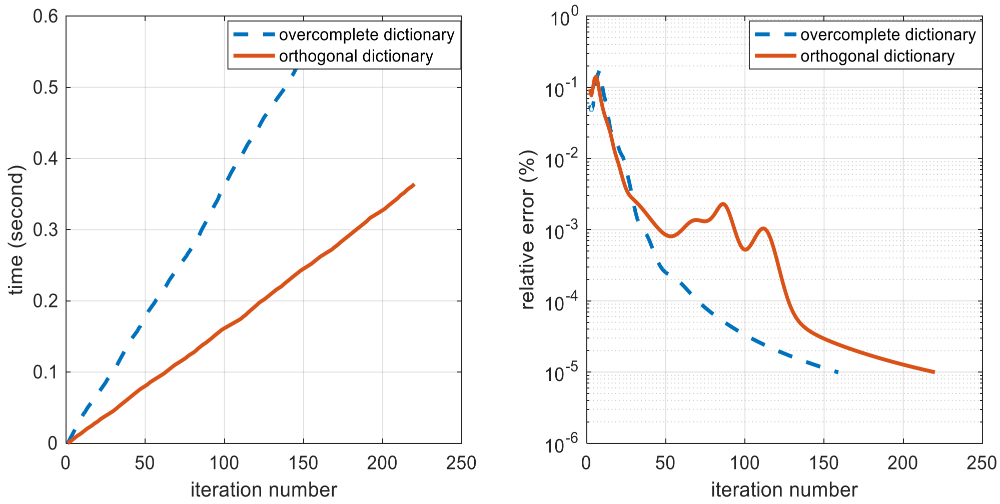

Figure 7. Comparisons of runtime and convergence performances of the proposed algorithm with orthogonal dictionary (size is 100) and overcomplete dictionary (size is 150).

6. Conclusions

In this paper we proposed a downlink channel estimation algorithm based on overcomplete dictionary and variational Bayesian inference. We converted the complex system model to a real model and exploited the correlation of angular channel sparsity in adjacent timeslots. In the algorithm we divided the timeslots into groups and made use of the channel support information of the previous timeslot to the channel estimation in the current timeslot within each group. The sparsity correlation and Bayesian Cramér–Rao bound for the MSE of channel estimation was analyzed. Compared with other recovery algorithms, such as WSP, IRLS, WIRLS, l_1 min, W-l_1 min and COSAMP, our proposed algorithm with overcomplete dictionary had a relatively better performance. Moderate overcomplete dictionary can improve the MSE performance of channel estimation to balance the computational complexity and performance gain.

Author Contributions: Conceptualization and methodology, W.L.; validation, X.W., S.P. and L.Z.; formal analysis, W.L.; writing—original draft preparation, W.L.; supervision, Y.W.

Funding: This work was supported in part by the National Science Foundation of China (No.61601509 and 61601334), the China Postdoctoral Science Foundation Grant (No.2016M603045 and 2018M632889) and the self-determined research funds of CCNU(CCNU18QN007) from the colleges basic research and operation of MOE.

Acknowledgments: Thank you to the reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Proof of angle change with UT movement.

According to the cosine law, we have

$$\cos\theta = \frac{d_1^2 + d_{LoS}^2 - (d_{NLoS} - d_1)^2}{2d_1 d_{LoS}}, \tag{A1}$$

$$\cos(\theta \pm \Delta_\theta) = \frac{d_1^2 + d_{LoS}^2 - (d_{NLoS} \pm d_\Delta - d_1)^2}{2d_1 d_{LoS}}. \tag{A2}$$

Then we can get

$$\theta \pm \Delta_\theta = \arccos\left(\frac{d_1^2 + d_{LoS}^2 - (d_{NLoS} - d_1)^2 - d_\Delta^2 \mp 2d_\Delta (d_{NLoS} - d_1)}{2d_1 d_{LoS}} \right). \tag{A3}$$

Since d_Δ^2 is very small compared with d_1 and d_{LoS}, by the first-order approximation we have

$$\begin{aligned} \theta \pm \Delta_\theta &\approx \arccos\left(\tfrac{d_1^2 + d_{LoS}^2 - (d_{NLoS} - d_1)^2 \mp 2d_\Delta (d_{NLoS} - d_1)}{2d_1 d_{LoS}} \right) \\ &\approx arc\cos\left(\tfrac{d_1^2 + d_{LoS}^2 - (d_{NLoS} - d_1)^2}{2d_1 d_{LoS}} \right) \pm \tfrac{2d_\Delta (d_{NLoS} - d_1)}{2d_1 d_{LoS}} \tfrac{1}{\sqrt{1-\cos^2\theta}} \quad . \\ &= \theta \pm \tfrac{2d_\Delta (d_{NLoS} - d_1)}{2d_1 d_{LoS}} \tfrac{1}{\sqrt{1-\cos^2\theta}} \end{aligned} \tag{A4}$$

Then we have

$$\Delta_\theta \approx \frac{2d_\Delta (d_{NLoS} - d_1)}{2d_1 d_{LoS}} \frac{1}{\sqrt{1 - \cos^2\theta}}. \tag{A5}$$

□

Appendix B

Proof of Proposition 1.

Let $\mathbf{z} \triangleq \left\{ \overline{h}, \sigma \right\}$, the we have

$$\mathbb{E}_{\mathbf{z}}\left\{ (\mathbf{z} - \hat{\mathbf{z}})(\mathbf{z} - \hat{\mathbf{z}})^T \right\} \geq \mathbf{J}^{-1}. \tag{A6}$$

Since $\overline{h}, \sigma$ are independent, the Fisher information matrix **J** is block diagonal, and can be presented as

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_{\overline{\mathbf{h}},\overline{\mathbf{h}}} & 0 \\ 0 & \mathbf{J}_{\sigma,\sigma} \end{bmatrix}. \tag{A7}$$

Then the inverse of matrix **J** is

$$\mathbf{J}^{-1} = \begin{bmatrix} {\mathbf{J}_{\overline{\mathbf{h}},\overline{\mathbf{h}}}}^{-1} & 0 \\ 0 & {\mathbf{J}_{\sigma,\sigma}}^{-1} \end{bmatrix}. \tag{A8}$$

Because $p(\overline{\boldsymbol{y}},\boldsymbol{z}) = p\left(\overline{\boldsymbol{y}}\middle|\boldsymbol{z}\right)p\left(\overline{\boldsymbol{h}}\middle|\boldsymbol{\alpha}\right)p\left(\boldsymbol{\alpha}\middle|\boldsymbol{b}\right)p(\boldsymbol{b})p(\sigma)$, we have

$$\begin{aligned}\mathbf{J} &= \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\overline{\mathbf{y}},\mathbf{z})}{\partial z_i \partial z_j}\right\} \\ &= \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\overline{\mathbf{y}}|\mathbf{z})}{\partial z_i \partial z_j}\right\} + \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\overline{\mathbf{h}}|\boldsymbol{\alpha})}{\partial z_i \partial z_j}\right\} + \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\boldsymbol{\alpha}|\mathbf{b})}{\partial z_i \partial z_j}\right\} + \\ &\mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\mathbf{b})}{\partial z_i \partial z_j}\right\} + \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\sigma)}{\partial z_i \partial z_j}\right\}\end{aligned} \tag{A9}$$

Since we mainly focus on the MSE of $\overline{\boldsymbol{h}}$, we only need to analyze $\mathbf{J}_{\overline{\mathbf{h}},\overline{\mathbf{h}}}$. We discuss the above formula part by part as follows:

1) Let $\mathbf{J}_{\overline{\mathbf{h}},\overline{\mathbf{h}}}(\overline{\mathbf{y}}) = \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\overline{\mathbf{y}}|\mathbf{z})}{\partial z_i \partial z_j}\right\}$, according to the Bayesian model in Figure 1, we have

$$p\left(\overline{\boldsymbol{y}}\middle|\boldsymbol{z}\right) \sim Normal(\overline{\mathbf{y}}|\overline{\mathbf{A}}\overline{\mathbf{h}}, \sigma\mathbf{I})\text{then}\mathbf{J}_{\overline{\mathbf{h}},\overline{\mathbf{h}}}(\overline{\mathbf{y}}) = \mathbb{E}_{\mathbf{z}}\{\frac{\overline{\mathbf{A}}^T\overline{\mathbf{A}}}{\sigma}\} = \frac{\overline{\mathbf{A}}^T\overline{\mathbf{A}}}{\sigma}.$$

2) Let $\mathbf{J}_{\overline{\mathbf{h}},\overline{\mathbf{h}}}(\overline{\mathbf{h}}) = \mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\overline{\mathbf{h}}|\boldsymbol{\alpha})}{\partial z_i \partial z_j}\right\}$, and we have $P\left(\overline{\mathbf{h}}\middle|\boldsymbol{\alpha}\right) = \prod_{i=1}^{2N} Normal\left(\overline{h}_i\middle|0,\alpha_i\right)$, then we get $\mathbf{J}_{\overline{\mathbf{h}},\overline{\mathbf{h}}}(\overline{\mathbf{h}}) = \mathbb{E}_{\mathbf{z}}\left\{\frac{1}{\alpha_i}\right\}$.

3) Because $\mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\boldsymbol{\alpha}|\mathbf{b})}{\partial z_i \partial z_j}\right\}$, $\mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\mathbf{b})}{\partial z_i \partial z_j}\right\}$ and $\mathbb{E}_{\mathbf{z}}\left\{-\frac{\partial^2 \log p(\sigma)}{\partial z_i \partial z_j}\right\}$ are independent with $\overline{\boldsymbol{h}}$, they are all 0. Then in summary, we get $\mathbf{J}_{\overline{\mathbf{h}},\overline{\mathbf{h}}} = diag(\mathbb{E}\left(\frac{1}{\alpha_i}\right)) + \frac{1}{\sigma^2}\overline{\mathbf{A}}^T\overline{\mathbf{A}}$.

Since the priori support set information is used in our proposed algorithm, a three-layer model is constructed for the elements belonging to the priori support set, and a two-layer model is used for the elements not belonging to the priori support set, so $\mathbb{E}_Z\left\{\frac{1}{\alpha_i}\right\}$ has different expressions for the two cases. $\mathbb{E}_Z\left\{\frac{1}{\alpha_i}\right\}$ in the two cases are discussed as follows:

1) When i belongs to the priori support set, according to the three-layer graph model we have

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{2N} Gamma\left(\alpha_i\middle|a,b_i\right), \tag{A10}$$

$$p(b_i) = \text{Gamma}\left(b_i\middle|c,d_i\right) = \Gamma(c)^{-1}d_i^c b_i^{c-1}e^{-d_i b_i}. \tag{A11}$$

Then we get

$$\begin{aligned}p(\alpha_i) &= \int_0^\infty p\left(\alpha_i\middle|b_i\right)p(b_i)db_i \\ &= \int_0^\infty \Gamma(a)^{-1}b_i^a\alpha_i^{a-1}e^{-b_i\alpha_i}\Gamma(c)^{-1}d_i^c b_i^{c-1}e^{-d_i b_i}db_i \\ &= \Gamma(a)^{-1}\Gamma(c)^{-1}\alpha_i^{a-1}d_i^c\frac{\Gamma(a+c)}{(a_i+d_i)^{a+c}}\end{aligned}. \tag{A12}$$

Accordingly, we have

$$\begin{aligned}\mathbb{E}\left\{\frac{1}{\alpha_i}\right\} &= \int_0^\infty \frac{1}{\alpha_i}\Gamma(a)^{-1}\Gamma(c)^{-1}\alpha_i^{a-1}d_i^c\frac{\Gamma(a+c)}{(a_i+d_i)^{a+c}}d\alpha_i \\ &= \int_0^\infty \frac{1}{\alpha_i}\frac{\Gamma(a+c)}{\Gamma(a)\Gamma(c)}\left(\frac{\alpha_i}{d_1}\right)^{a-1}\left(\frac{\alpha_i}{d_1}+1\right)^{-a-c}d\frac{\alpha_i}{d_1}\end{aligned}, \tag{A13}$$

where $\frac{\Gamma(a+c)}{\Gamma(a)\Gamma(c)}\left(\frac{\alpha_i}{d_i}\right)^{a-1}\left(\frac{\alpha_i}{d_i}+1\right)^{-a-c}$ satisfies the probability density function of Beta prime distribution. According to the properties of the Beta prime distribution, when $-a < -1 < c$, we have

$$\mathbb{E}\left\{\left(\overline{\mathbf{h}}-\hat{\overline{\mathbf{h}}}\right)^{H}\left(\overline{\mathbf{h}}-\hat{\overline{\mathbf{h}}}\right)\right\} \begin{array}{l} \geq |S|\cdot\frac{1}{|S|}\sum\limits_{i\in S}\frac{1}{\frac{1+c}{a\min(d)}++\frac{\lambda_i}{\sigma}}+(N-|S|)\cdot\frac{1}{(N-|S|)}\sum\limits_{i\notin S}\frac{1}{\frac{\max(b)}{a}+\frac{\lambda_i}{\sigma}} \\ \to |S|\frac{a\min(\mathbf{d})}{1+c}\left(1-\frac{F(snr_1,\beta)}{4\beta snr_1}\right)+(N-|S|)\frac{a}{\max(\mathbf{b})}\left(1-\frac{F(snr_2,\beta)}{4\beta snr_2}\right)\end{array}. \quad \text{(A14)}$$

2) When i does not belong to the priori support set, according to the high-order moment properties for the general gamma distribution, we have

$$\mathbb{E}\left\{\frac{1}{\alpha_i}\right\}=\frac{b_i}{a}. \quad \text{(A15)}$$

Then in summary, we have

$$\mathbb{E}\left\{\|\overline{\mathbf{h}'}-\overline{\mathbf{h}}\|^{2}\right\}\geq tr\left(\left(diag(\mathbb{E}\left(\frac{1}{\alpha_i}\right))+\frac{1}{\sigma^2}\overline{\mathbf{A}}^{T}\overline{\mathbf{A}}\right)^{-1}\right)=\sum_{i\in S}\frac{1}{\frac{1+c}{ad_i}+\frac{\lambda_i}{\sigma}}+\sum_{i\notin S}\frac{1}{\frac{b_i}{a}+\frac{\lambda_i}{\sigma}}, \quad \text{(A16)}$$

where S is the diagnosed support set, λ_i is the eigenvalues of $\overline{\mathbf{A}}^{\mathrm{T}}\overline{\mathbf{A}}$,and $\overline{\mathbf{A}}^{\mathrm{T}}\overline{\mathbf{A}}\in\mathbb{R}^{\mathrm{2M\times 2M}}$.

When overcomplete dictionary is as $\boldsymbol{D}^{d}=\left\{\frac{1}{\sqrt{N}}e^{-j\frac{2\pi}{M}kn}\right\}_{n,k}$, $k\in\{1,\cdots,M\}$, $n\in\{1,\cdots,N\}$, and $\mathbf{A}$ is Gaussian random matrix with each element is mean 0 and variance $\frac{1}{T_d}$, then $\mathbf{AD}^{\mathrm{d}}$ is complex Gaussian random matrix. Then $\overline{\mathbf{A}}$ is Gaussian random matrix with mean 0 and variance $\frac{\rho^d}{2T_d}$.

According to the random matrix theory, for $N\times K$ dimensional random matrix $\mathbf{H}$ with each element is independent and is variable with mean 0 and variance $1/N$, when $K,N\to\infty$ and $\frac{K}{N}\to\beta$,then the empirical distribution of eigenvalues of $\mathbf{H^{T}H}$ converges almost surely as $f_\beta(x)=\left(1-\frac{1}{\beta}\right)^{+}\delta(x)+\frac{\sqrt{(x-a)^{+}(b-x)^{+}}}{2\pi\beta x}$, where $(x)^{+}=\max(0,x)$, $a=\left(1-\sqrt{\beta}\right)^{2}$, $b=\left(1+\sqrt{\beta}\right)^{2}$.

Since $\overline{\mathbf{A}}\in\mathbb{R}^{2T_d\times 2M}$, and its element is Gaussian random variable with mean 0 and variance $\frac{\rho^d}{2T_d}$ By applying the above results for the empirical distribution of eigenvalues of $\mathbf{H^{T}H}$, when $T_d,M\to\infty$ and $\frac{T_d}{M}=\beta$, the empirical distribution of eigenvalues λ of $\overline{\mathbf{A}}^{\mathbf{T}}\overline{\mathbf{A}}$ converges almost surely as

$$f_\beta(\lambda)=\left(1-\frac{1}{\beta}\right)^{+}\delta(\lambda)+\frac{\sqrt{(\lambda-a\prime)^{+}(b\prime-\lambda)^{+}}}{2\pi\beta\lambda\sqrt{\rho^d}} \quad \text{(A17)}$$

where $a\prime=\sqrt{\rho^d}\left(1-\sqrt{\beta}\right)^{2}$, $b\prime=\sqrt{\rho^d}\left(1+\sqrt{\beta}\right)^{2}$. When $s,M\to\infty$ and $\frac{s}{M}=\mu$, we have

$$\mathbb{E}\left\{\left(\overline{\mathbf{h}}-\hat{\overline{\mathbf{h}}}\right)^{H}\left(\overline{\mathbf{h}}-\hat{\overline{\mathbf{h}}}\right)\right\} \begin{array}{l} \geq |S|\cdot\frac{1}{|S|}\sum\limits_{i\in S}\frac{1}{\frac{1+c}{a\min(d)}++\frac{\lambda_i}{\sigma}}+(N-|S|)\cdot\frac{1}{(N-|S|)}\sum\limits_{i\notin S}\frac{1}{\frac{\max(b)}{a}+\frac{\lambda_i}{\sigma}} \\ \to |S|\frac{a\min(\mathbf{d})}{1+c}\left(1-\frac{F(snr_1,\beta)}{4\beta snr_1}\right)+(N-|S|)\frac{a}{\max(\mathbf{b})}\left(1-\frac{F(snr_2,\beta)}{4\beta snr_2}\right)\end{array}, \quad \text{(A18)}$$

where $snr_1=\frac{a\min(\mathbf{d})}{(1+c)\sigma}$, $snr_2=\frac{a}{\sigma\max(\mathbf{b})}$ and $F(x,z)=\left(\sqrt{x\left(1+\sqrt{z}\right)^{2}+1}-\sqrt{x\left(1-\sqrt{z}\right)^{2}+1}\right)^{2}$.

Then the proofs are complete. □

References

1. Lu, L.; Li, G.Y.; Swindlehurst, A.L.; Ashikhmin, A.; Zhang, R. An overview of Massive MIMO: Benefits and challenges. *IEEE J. Sel. Top. Signal Process.* **2014**, *8*, 742–758. [CrossRef]
2. Rao, X.; Lau, V.K.N. Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems. *IEEE Trans. Signal Process.* **2014**, *12*, 3261–3271.
3. Vaswani, N.; Lu, W. Modified-CS: Modifying compressive sensing for problems with partially known support. *IEEE Trans. Signal Process.* **2010**, *9*, 4595–4607. [CrossRef]
4. Borries, R.V.; Miosso, C.; Potes, C. Compressed sensing using priori information. In Proceedings of the 2nd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive, St. Thomas, VI, USA, 12–14 December 2007; pp. 121–124.
5. Tseng, C.C.; Wu, J.Y.; Lee, T.S. Enhanced compressive downlink CSI Recovery for FDD Massive MIMO systems using weighted Block l_1 minimization. *IEEE Trans. Commun.* **2016**, *3*, 1055–1066. [CrossRef]
6. Lu, W.; Wang, Y.; Fang, Q.; Peng, S. Downlink compressive channel estimation with support diagnosis in FDD massive MIMO. *J. Wirel. Commun. Netw.* **2018**, *115*, 1–12. [CrossRef]
7. Masood, M.; Afify, L.H.; Al-Naffouri, T.Y. Efficient coordinated recovery of sparse channels in massive MIMO. *IEEE Trans. Signal Process.* **2015**, *1*, 104–118. [CrossRef]
8. Cheng, X.; Sun, J.; Li, S. Channel estimation for FDD multi-user massive MIMO: A variational Bayesian inference-based approach. *IEEE Trans. Wirel. Commun.* **2017**, *11*, 7590–7602. [CrossRef]
9. Dai, J.; Liu, A.; Lau, V.K.N. FDD massive MIMO channel estimation with arbitrary 2D-Array Geometry. *IEEE Trans. Signal Process.* **2018**, *10*, 2584–2599. [CrossRef]
10. Xie, H.; Gao, F.; Jin, S.; Fang, J.; Liang, Y. Channel Estimation for TDD/FDD Massive MIMO Systems with Channel Covariance computing. *IEEE Trans. Wirel. Commun.* **2018**, *6*, 4206–4218. [CrossRef]
11. Shen, W.; Dai, L.; Shi, Y.; Shim, B.; Wang, Z. Joint Channel Training and Feedback for FDD Massive MIMO Systems. *IEEE Trans. Veh. Technol.* **2016**, *10*, 8762–8767. [CrossRef]
12. Tauböck, G.; Hlawatsch, F.; Eiwen, D.; Rauhut, H. Compressive Estimation of Doubly Selective Channels in Multicarrier Systems: Leakage Effects and Sparsity-Enhancing Processing. *IEEE J. Sel. Top. Signal Process.* **2010**, *2*, 255–271. [CrossRef]
13. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
14. Christopher, M. Bishop. Bayesian Linear Regression. In *Pattern Recognition and Machine Learning*; Springer-Verlag: Heidelberg, Germany, 2006; pp. 152–160.
15. Fang, J.; Shen, Y.; Li, F.; Li, H.; Chen, Z. Support knowledge-aided sparse Bayesian learning for compressed sensing. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 19–24 April 2015; pp. 3786–3790.
16. Ji, S.; Xue, Y.; Carin, L. Bayesian compressive sensing. *IEEE Trans. Signal Process.* **2008**, *6*, 2346–2356. [CrossRef]
17. Lu, W.; Wang, Y.; Fang, Q.; Peng, S. Compressive Channel Estimation Based on Weighted IRLS in FDD Massive MIMO. *Wirel. Pers. Commun.* **2018**, *2*, 1–10. [CrossRef]
18. Candes, E.J.; Romberg, J.K.; Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **2006**, *8*, 1207–1223. [CrossRef]