Efficient Channel State Information Acquisition in Massive MIMO Systems using Non-Orthogonal Pilots

Paul Ferrand, Alexis Decurninge, Maxime Guillaud, Luis G. Ordóñez
Mathematical and Algorithmic Sciences Laboratory, Huawei Technologies
France Research Center, 20 quai du Point du Jour, 92100 Boulogne Billancourt, France

***Abstract*—The objective of this article is to review and benchmark strategies to acquire channel state information (CSI) in time-division duplex (TDD) Massive MIMO systems. In particular, we consider the use of statistical CSI at the base stations (BSs), together with non-orthogonal pilot sequences. Such techniques can theoretically reduce the amount of spectral resources dedicated to channel sounding, thanks to the increased pilot reuse that they allow. We review their application, and present a holistic evaluation of their practical implementation and limitations, including factors that are often neglected in theoretical works, including covariance estimation and tracking, non-orthogonal pilot sequences, signaling associated with dynamically allocated pilot sequences, realistic channel model (spatial covariance, stationarity), and imperfect statistical CSI. We present a possible design addressing these various problems with reasonable complexity, and benchmark the achieved performance. Our results show that when the cost of CSI acquisition is properly accounted for, the tracking of statistical CSI together with the use of non-orthogonal pilots enables to reach a higher effective spectral efficiency than what can be achieved with orthogonal pilot sequences.**

I. Introduction

Massive MIMO (Multiple-Input Multiple-Output) [1] is considered to be one of the key technologies for realizing the performance targets expected from future 5G wireless systems [2]. Reaping all the benefits of Massive MIMO systems shown e.g. in [3] requires acquiring accurate channel state information (CSI) at the base stations (BSs). However, the large number of antenna elements in Massive MIMO BSs makes the design of the CSI acquisition strategy particularly challenging.

In the presence of i.i.d. channels, it is well known that the length of the pilot sequences should be at least as large as the total number of transmit antennas in the system [4], in order to avoid the effect known as *pilot contamination* [5]. Depending on the coherence time of the channel, the transmission of long training sequences instead of data-bearing symbols can represent a significant loss in spectral efficiency. This has led to considering pilot reuse within the network, first under the assumption of random pairing of users and pilots within each cell or cluster, as in [6], [7].

In the context of Massive MIMO systems, channels exhibit a large degree of spatial correlation [8]. Dense antenna arrays make the signals received at the BS more spatially correlated, to the point of resulting in a rank deficient spatial correlation matrix [9]. This correlation can potentially help to reduce the required training overhead, and is likely to be a key driver in the design of Massive MIMO systems [10]. The problem of pilot design for MIMO correlated channels has been widely studied in the point-to-point case, for example in [11]–[13]. Extensions to the case of imperfect statistical information have been proposed in [14].

As a first approach, statistical information to mitigate pilot contamination has been studied mostly under the assumption that a fixed set of orthogonal pilot sequences is reused, each one being assigned to one or several users across one or several cells on the basis of statistical CSI information. In addition to reuse across cells, such approaches enable some degree of pilot reuse *within a cell* (fractional pilot reuse), leading to a lower overhead associated to the reference symbols. In [9], [15], [16], this is enabled by assuming that per-user channel covariance information is available at the BS, while statistical CSI is limited to per-user path loss information in [17]. Indeed, (perfect) covariance knowledge has been shown to potentially remove the theoretical limit on pilot contamination, as in [18] where an asymptotically unbounded SINR is obtained under a fairly loose assumption on the linear independance of per-user covariance matrices.

Conversely to the reuse of orthogonal pilot sequences, in [19] it was proposed to assign arbitrary—possibly neither orthogonal nor identical—pilot sequences to users in a system on the basis of known per-user channel covariance information. This approach allows for a more flexible trade-off between the number of users per pilot resource, and the level of pilot contamination. In frequency-division duplex (FDD) systems, downlink CSI can not be simply obtained by reciprocity [20] at the BS; in that scenario, the large number of BS antennas makes the use of orthogonal pilot sequences especially unwelcome. An approach to design non-orthogonal pilots for downlink channel estimation, more adapted to systems where channel reciprocity does not hold, e.g. FDD-based systems, is proposed in [21].

The main question that this work thus aims to answer is what—and how much—can be gained by tracking statistical CSI at the BS. To this end, we introduce and benchmark a particular architecture for the CSI acquisition strategy of

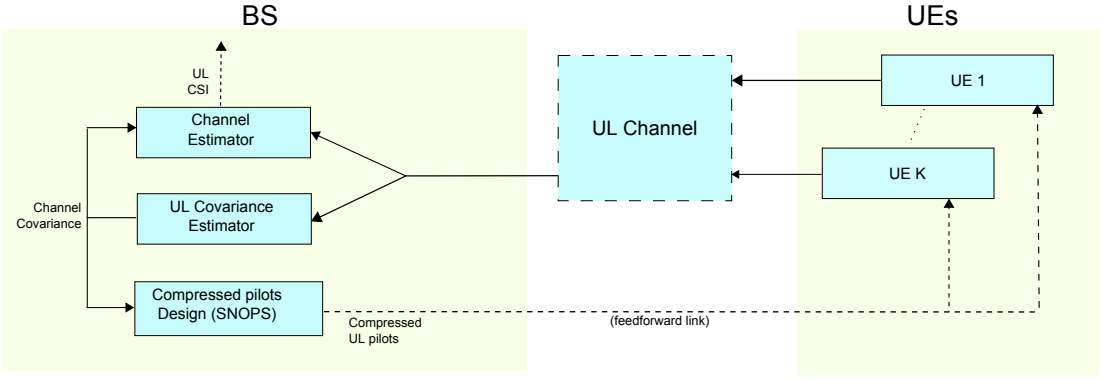

Fig. 1. Block diagram of a Covariance-Based CSI acquisition system.

future cellular networks operating in time-division duplex (TDD) mode; the considered architecture is centered around the tracking of statistical CSI at the BS, and seeks to optimize the amount of spectral resources dedicated to CSI acquisition.

Our objective here is to outline a possible holistic design taking into account a practical strategy for statistical CSI estimation, design and allocation of non-orthogonal pilots, computationally feasible vector quantization approaches for the feedforward links, together with low-complexity (linear) multi-user precoding and equalization approaches.

II. Covariance-based CSI Acquistion Framework

Let us focus on CSI acquisition in the uplink, since it is expected that few downlink pilots will be used in Massive MIMO systems. If uplink CSI is available at the BS, downlink CSI can be obtained at the BS via reciprocity and enable the use of multiuser precoding. Furthermore, downlink CSI is considered unnecessary at the user equipment (UE) [22]. The amount of uplink reference symbols is thus expected to be a bottleneck for TDD Massive MIMO systems. We first consider the problem of uplink CSI estimation under the assumption that the corresponding statistical CSI is available, in the form of per-user spatial covariance information at the BS side. We then present a scheme to acquire and track this spatial covariance information. For the sake of clarity of the exposition, we first consider a single-cell system; extension to multiple cells is finally discussed in Section V.

Figure 1 depicts the various functionalities of the considered covariance-based CSI acquisition system. The role of the various blocks is as follows:

- The K UEs concurrently transmit their respective pilot sequences on the uplink. The pilot sequences are determined by the BS.
- The channel estimator merges the available statistical CSI and the information from the received reference symbols to compute an estimate of the instantaneous uplink CSI. Downlink CSI can therefore also be obtained at the BS via reciprocity.
- The pilot sequences are periodically computed or allocated to the users, based on the available statistical CSI. This allows for shorter (non-orthogonal) uplink training sequences.
- The UEs are informed of the chosen pilot sequences through proper quantization and signaling through a feedforward link.
- A covariance estimation algorithm keeps track of statistical CSI (in the form of per-user spatial covariance matrices) at the BS, based on the uplink training signals.

A. Channel Model

Let $\mathbf{H}_k \in \mathbb{C}^{M\times N}$ denote the channel matrix containing the narrowband channel coefficients between the N antennas of the k-th UE and the M BS antennas and let $\mathbf{H}_k$ be modeled as

$$\mathbf{H}_k = \mathbf{R}_k^{\frac{1}{2}}\mathbf{F}_k, \qquad k = 1,\dots,K. \tag{1}$$

where $\mathbf{R}_k = \mathbb{E}[\mathbf{H}_k\mathbf{H}_k^H]/N \in \mathbb{C}^{M\times M}$ is the rank-r_k spatial covariance matrix at the BS and matrix $\mathbf{F}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{r_k}) \in \mathbb{C}^{r_k\times N}$ represents the fast-fading process.

Observe that the model in (1) implicitly assumes that the N channels corresponding to the N antennas of the k-th user undergo the same BS-side correlation. Furthermore, we assume that the channel model in (1) is locally stationary in time, i.e., the spatial covariance matrices $\{\mathbf{R}_k\}$ remain constant, while the fast-fading processes $\{\mathbf{F}_k\}$, independent across UEs, undergo several independent fades [23].

B. Channel Estimation

The training-based CSI acquisition strategy can be summarized as follows. Let $\mathbf{P}_k = [\boldsymbol{p}_k(1),\dots,\boldsymbol{p}_k(L)] \in \mathbb{C}^{N\times L}$ gather the N pilot sequences of length L transmitted by the N antennas of user k. Letting all K users concurrently transmit their respective pilot sequences, the signal received by the BS at the ℓ-th channel use is

$$\boldsymbol{y}(\ell) = \sum_{k=1}^{K}\mathbf{H}_k\boldsymbol{p}_k(\ell) + \boldsymbol{n}(\ell), \qquad \ell = 1,\dots,L \tag{2}$$

where $\boldsymbol{n}(\ell) = [n_1(\ell),\dots,n_M(\ell)]^T \in \mathbb{C}^M$ denotes the additive white Gaussian noise with i.i.d. components $n_i(\ell) \sim \mathcal{CN}(0,\sigma^2)$. Grouping the received signal for the L channel

uses as $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(L)] \in \mathbb{C}^{M \times L}$, the system model in (2) can be more compactly written as

$$\mathbf{Y} = \mathbf{H}\mathbf{P} + \mathbf{N} \tag{3}$$

where $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_K]$ is the column concatenation of the channel matrices from the K UEs, $\mathbf{P} = [\mathbf{P}_1^T, \ldots, \mathbf{P}_K^T]^T \in \mathbb{C}^{KN \times L}$ is the matrix containing all the training sequences, and $\mathbf{N} = [\mathbf{n}(1), \ldots, \mathbf{n}(L)] \in \mathbb{C}^{M \times L}$.

The BS adopts a linear minimum mean-square error (LMMSE) channel estimation strategy, which exploits the knowledge on the users channel covariances $\{\mathbf{R}_k\}$ and the noise power σ^2. Let us define $\boldsymbol{y} \triangleq \mathrm{vec}(\mathbf{Y})$, $\boldsymbol{h} \triangleq \mathrm{vec}(\mathbf{H})$ and $\boldsymbol{n} \triangleq \mathrm{vec}(\mathbf{N})$ and rewrite the received signal in (3) as

$$\boldsymbol{y} = \tilde{\mathbf{P}}\boldsymbol{h} + \mathbf{n} \tag{4}$$

where $\tilde{\mathbf{P}} = (\mathbf{P}^T \otimes \mathbf{I}_M)$. Combining (1) with (4) yields

$$\boldsymbol{y} = \tilde{\mathbf{P}}\tilde{\mathbf{R}}^{\frac{1}{2}}\boldsymbol{f} + \boldsymbol{n} \tag{5}$$

where $\tilde{\mathbf{R}}^{\frac{1}{2}} = \mathrm{diag}\big(\mathbf{I}_N \otimes \mathbf{R}_1^{\frac{1}{2}}, \ldots, \mathbf{I}_N \otimes \mathbf{R}_K^{\frac{1}{2}}\big)$, and $\boldsymbol{f} \triangleq \mathrm{vec}(\mathbf{F}) \in \mathbb{C}^{rN}$ is the vector concatenation of the fast-fading process coefficients, with $r = \sum_{k=1}^{K} r_k$. Then, the LMMSE estimate of $\boldsymbol{h}$ is given by

$$\hat{\boldsymbol{h}} = \tilde{\mathbf{R}}\tilde{\mathbf{P}}^H(\tilde{\mathbf{P}}\tilde{\mathbf{R}}\tilde{\mathbf{P}}^H + \sigma^2\boldsymbol{I}_{ML})^{-1}\boldsymbol{y}. \tag{6}$$

C. Non-Orthogonal Pilot Allocation

Various approaches have been proposed to allocate pilots (in our notations, choose $\mathbf{P}$) on the basis of statistical CSI. For instance, the authors of [9] and [15] seek to cluster the users based on their covariance. Pilots are reused among the users having sufficiently orthogonal channel subspaces; here, the non-orthogonality takes the form of identical reuse of pilots across certain users inside a cell or between cells. More recently, the necessity of covariance matrices spanning orthogonal subspaces was relaxed to the milder condition of linear independence between the covariance matrices [18]. In [19], we proposed a generic approach for arbitrary covariance matrices and completely arbitrary pilot sequences with varying degrees of orthogonality. Whereas previous work [9], [18] focused on removing the theoretical limitation of pilot contamination in Massive MIMO, our goal is to use the covariance side-information to reduce the size of the pilots and therefore increase the overall spectral efficiency of the system. Take an extreme case as an example, and assume that the covariance matrices of all users are mutually orthogonal. In that case, similarly to [15], we can separate the signals from all users in the spatial domain and therefore only require a single pilot symbol. On the other hand, if all covariance matrices span the same subspace, then we have to use orthogonal pilots to ensure the identifiability of the channels across all users—see [19, Section III] We therefore fall back to the more classical schemes. The approach that we follow here provides a middle ground between these two extremes cases. We shortly describe it now; we refer the interested reader to [19] for more details on the analysis and implementation.

From (6) we can write the covariance of the estimation error on $\boldsymbol{h}$ as

$$\begin{aligned}\mathbf{C}_e &= \mathbb{E}\big[(\hat{\boldsymbol{h}} - \boldsymbol{h})(\hat{\boldsymbol{h}} - \boldsymbol{h})^H\big] \\ &= \tilde{\mathbf{R}} - \tilde{\mathbf{R}}\tilde{\mathbf{P}}^H(\tilde{\mathbf{P}}\tilde{\mathbf{R}}\tilde{\mathbf{P}}^H + \sigma^2\boldsymbol{I}_{ML})^{-1}\tilde{\mathbf{P}}\tilde{\mathbf{R}}.\end{aligned} \tag{7}$$

Our aim is to minimize the length of the pilot sequences over all K users, while bounding the estimation error over all components of the channel vector $\boldsymbol{h}$. To ensure this last property, we can require that all the eigenvalues of $\mathbf{C}_e$ are below an error threshold $\varepsilon > 0$, which can be written as a generalized inequality on the cone of positive semi-definite matrices [24]. The constraint thus becomes $\mathbf{C}_e \preceq \varepsilon\mathbf{I}$, where $\mathbf{A} \preceq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is a positive semi-definite matrix. Meanwhile, assuming that $L \leq KN$, minimizing the length of the pilot sequences is equivalent to minimizing $\mathrm{rank}(\mathbf{X})$ where $\mathbf{X} = \mathbf{P}\mathbf{P}^H$. Overall, we can cast the non-orthogonal pilot allocation problem as:

$$\begin{aligned}\min_{\mathbf{X} \succeq 0} \quad & \mathrm{rank}(\mathbf{X}) \\ \text{s.t.} \quad & \mathbf{C}_e \preceq \varepsilon\mathbf{I}.\end{aligned} \tag{8}$$

The objective function is not smooth but it can be approximated by a smooth concave function of $\mathbf{X}$. On the other hand, the constraint on $\mathbf{C}_e$ can be transformed towards a convex constraint in $\mathbf{X}$:

$$\tilde{\mathbf{R}}^{\frac{H}{2}}(\mathbf{X} \otimes \mathbf{I}_M)\tilde{\mathbf{R}}^{\frac{1}{2}} \preceq (\varepsilon^{-1}\boldsymbol{\Lambda} - \mathbf{I}_{rN})\sigma^2 \tag{9}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues of $\tilde{\mathbf{R}}$. Thus, the original problem can be approximated by a semi-definite program, which be efficiently solved using numerical algorithms for convex optimization.

III. Covariance estimation

In the available literature, the analysis of covariance matrix estimation is usually separated from the pilot design question. This typically leads to two separate pilot design approaches (one for instantaneous CSI estimation, and one for covariance estimation) following different requirements—the practical consequence being that the use of two separate (and orthogonal) sets of pilots is required, as in [25].

In this section, we point out that the use of non-orthogonal pilots opens up new challenges in this area. Indeed, since many measurements are performed for CSI acquisition purposes, it is natural to use them in order to estimate the underlying covariance information. We propose therefore to use the pilot sequences used for channel estimation and extend them in order to obtain a reliable covariance estimation.

We assume local stationarity of the channels statistics. We can thus estimate this covariance according to T past measurements (where T depends on the covariance coherence time) based on previous sets of pilot sequences. Rewriting eq. (4) for each round of channel estimation indexed by $t = 1 \ldots T$, and letting $\mathbf{P}_t \in \mathbb{C}^{KN \times L_t}$ denote the set of length-L_t pilot sequences used for round t, the received signal

$\boldsymbol{y}_t \triangleq \text{vec}(\mathbf{Y}_t)$ is a projection of the concatenation of all channels $\boldsymbol{h}_t$:

$$\boldsymbol{y}_t = \tilde{\mathbf{P}}_t \boldsymbol{h}_t + \boldsymbol{n}_t \tag{10}$$

where $\tilde{\mathbf{P}}_t = (\mathbf{P}_t^T \otimes \mathbf{I}_N)$. Note that if the pilot matrix $\mathbf{P}_t$ is invertible—which is equivalent to the invertibility of $\tilde{\mathbf{P}}_t$)—the whole channel $\boldsymbol{h}_t$ is observed up to a linear transformation. A sufficient condition for this invertibility is that $\mathbf{P}_t$ is orthogonal and full-length, i.e. $L_t = KN$. On the other hand, if $\mathbf{P}_t$ is not full-length and $L_t < KN$, then only a projection of the channel is observed. These two cases lead to different approaches in covariance estimation.

A. Orthogonal (full-length) pilots

If orthogonal pilots are in use, the covariance estimation and tracking can simply be performed without additional signaling. Indeed, in that case each $\tilde{\mathbf{P}}_t$ is invertible and a natural estimator corresponds to the sample covariance estimator. It can be based e.g. on transformed measurements $\left(\tilde{\mathbf{P}}_t^{-1}\boldsymbol{y}_t\right)_{t=1\dots T}$.

B. Non-orthogonal pilots

In case of non-orthogonal pilots, if we assume that the same pilot matrix $\mathbf{P}$ of length $L < KN$ is used for T channel acquisitions, we will observe through $\boldsymbol{y}_t$ the projections of the concatenated channel onto the subspace spanned by the rows of $\tilde{\mathbf{P}}$. Any covariance estimator will thus be biased if the rows of $\tilde{\mathbf{P}}$ do not span the complete subspace of the non-zero eigenvectors of the true covariance matrix $\tilde{\mathbf{R}}$; this can be the case e.g. if $\mathbf{P}$ is calculated based on outdated or incorrectly estimated statistical CSI. In this case, extra pilot sequences are needed to obtain a reliable unbiased estimation and correct the error on the statistical CSI. We now introduce an approach to design such extra pilot sequences.

Let $\mathbf{Q}_t = (\boldsymbol{q}_{1,t}, \dots, \boldsymbol{q}_{J_t,t})$ denote this additional sequence of length J_t. The augmented t-th set of pilot sequences of length $L_t = J_t + L_t$ will be

$$\mathbf{S}_t = (\mathbf{P}_t, \mathbf{Q}_t). \tag{11}$$

In order to have a consistent estimation of $\mathbf{R}_1, \dots, \mathbf{R}_K$, it is sufficient for the columns of $\mathbf{S}_1, \dots \mathbf{S}_T$ to span the whole space $\mathbb{C}^{KN}$. For example, we could consider a simple example of design of $\mathbf{Q}_t$ where $J_t = 1$ and $\mathbf{Q}_t$ is a rotating Fourier vector

$$\mathbf{Q}_t = \left(1, e^{2i\pi\frac{(t-1)}{KN}}, \dots, e^{2i\pi(KN-1)\frac{(t-1)}{KN}}\right)^T \tag{12}$$

This results in a low impact on the pilot load while allowing for consistent covariance estimators.

C. Covariance Estimation: batch, online and tracking approaches

Note that the received signals $\boldsymbol{y}_t$ are zero-mean Gaussian random vectors of covariance $\tilde{\mathbf{P}}_t\tilde{\mathbf{R}}\tilde{\mathbf{P}}_t^H + \sigma^2\mathbf{I}_{NL_t}$ with $\tilde{\mathbf{R}}$ depending linearly on the covariances matrices $\mathbf{R}_1, \dots, \mathbf{R}_K$. A classical way to estimate the underlying parameters $\mathbf{R}_1, \dots, \mathbf{R}_K$ would then be to consider the maximum likelihood given the observation of a batch of samples $\boldsymbol{y}_1, \dots, \boldsymbol{y}_t$. However, the computation of the maximum likelihood estimate may be intractable for large T. An online algorithm approximating the maximum likelihood estimator is then of interest. It consists in iterating the estimation each time a new sample is observed using only this new sample and the previous estimate. However, such estimators are assuming that the true covariance is stationary while we consider only a local temporal stationarity of the covariance. We therefore suggest to adapt this online estimation in order to track the covariance variations over time in the spirit of [26] and [27] where online algorithms are developed to track the underlying subspace from the sequence of observations.

D. Imperfect Covariance Knowledge

In practice and through the estimators we described, we have imperfect knowledge of the covariance; we only have access to an estimate $\hat{\mathbf{R}}_k$ of $\mathbf{R}_k$. This covariance imperfection introduces a bias in the channel estimation; see [14]. We propose a model on the covariance estimate that allows to take the imperfection into account in the channel estimation. In order to avoid the positive definiteness constraint, we will consider that this imperfection can be modeled as the following additive noise for each user k on the square root of the covariance matrix:

$$\mathbf{R}_k^{\frac{1}{2}} = \hat{\mathbf{R}}_k^{\frac{1}{2}} + \mathbf{E}_k \tag{13}$$

with $\mathbf{E}_k$ a zero-mean random matrix that we will suppose independent from each channel random matrix $\mathbf{H}_1, \dots, \mathbf{H}_K$ for simplicity. In that case, the receive covariance of the random matrix $\mathbf{H}_k$ is

$$\frac{1}{N}\mathbb{E}[\mathbf{H}_k\mathbf{H}_k^H] = \hat{\mathbf{R}}_k + \mathbb{E}[\mathbf{E}_k\mathbf{E}_k^H]. \tag{14}$$

Imperfect covariance knowledge can then be seen as an other source of noise for the channel estimator. Assuming furthermore $\mathbb{E}[\mathbf{E}_k\mathbf{E}_k^H] = \sigma_R^2\mathbf{I}_M$, the LMMSE estimator (6) is now

$$\begin{aligned}\hat{\boldsymbol{h}} =&(\hat{\mathbf{R}} + \sigma_R^2\mathbf{I}_{KNM})\cdot \\ &\tilde{\mathbf{P}}^H\left(\tilde{\mathbf{P}}(\hat{\mathbf{R}} + \sigma_R^2\mathbf{I}_{KNM})\tilde{\mathbf{P}}^H + \sigma^2\boldsymbol{I}_{LM}\right)^{-1}\boldsymbol{y}.\end{aligned} \tag{15}$$

IV. An Example of a Practical System

Since the emphasis of this work is on CSI acquisition, we described in detail this aspect of the proposed architecture in the previous section. In order to evaluate the performance of our CSI acquisition scheme, we consider the single-cell multi-user system architecture pictured in Figure 2; we choose the uplink throughput as the performance benchmark to evaluate the acuracy and efficiency of the CSI estimation. The proposed architecture has a number of desirable properties, among which: (i) covariance estimation does not require dedicated training signals, and is based on the compressed pilots used for instantaneous CSI acquisition; and (ii) UL precoders—and potentially DL precoders using reciprocity—are centrally computed. This approach thus avoids the suboptimality associated with distributed decisions.

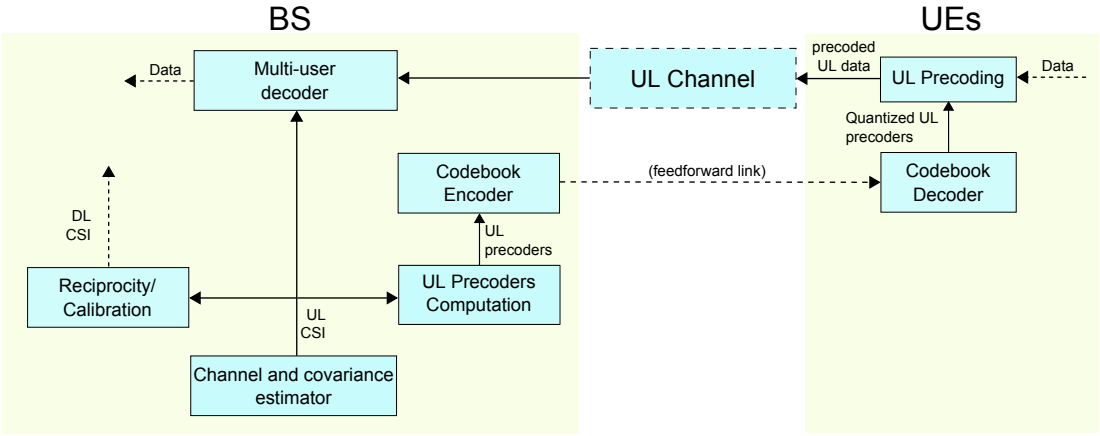

Fig. 2. Functional block diagram of the single-cell system considered in the simulation.

We consider a basic TDD-LTE system with a Orthogonal Frequency Division Multiplexing (OFDM) scheme in frequency—see [28] and Table I for the specific numerology we use. We assume that all the users share the same time-frequency resources and are thus multiplexed in the spatial domain. To obtain a fair assessment of the performance of a CSI acquisition scheme, a key question in the performance evaluation is related to the granularity of pilot signals in the frequency domain. Let the total system bandwidth be split into N_{sc} frequency subcarriers, and let N_{ts} indicate the number of time slots constituting a frame. The total number of time-frequency resources in a frame is thus $N_{\mathrm{sc}} \cdot N_{\mathrm{ts}}$. The channel is assumed constant within a resource block (RB), therefore CSI estimates only need to be acquired once within each resource block. Pilots from all users in the frame will then consume $N \cdot L$ time-frequency resources per RB, leading to an efficiency factor

$$\eta = 1 - \frac{N \cdot L}{N_{\mathrm{sc}} \cdot N_{\mathrm{ts}}}. \tag{16}$$

Using the channel estimates, the BS can then compute both uplink precoders and feed them forward to the users, and the multi-user decoders to separate the received symbols from each users. The number of resources per RB $N_{\mathrm{sc}} \cdot N_{\mathrm{ts}}$ is therefore a key parameter to consider in practical situations. It is especially important for the decoder, and thus we focus more on the *decoding* bandwidth than the *precoding* bandwidth that has less of an impact on the system performance. The decoding bandwidth is then the frequency bandwidth over which the multi-user decoder at the BS assumes constant CSI. Decreasing the decoding bandwidth or the number of time slots in a frame should improve the system performance since the system will obtain better channel estimates. However, this will decrease the overall efficiency for a fixed number of users and user antennas. On the other hand, for channels with larger coherence bandwidth or channels with larger temporal coherence we can afford a larger $N_{\mathrm{sc}} \cdot N_{\mathrm{ts}}$ factor. For these, short pilots will provide less significant gains.

A. Vector quantization on the feedforward link

Quantization is in heavy use in the considered architecture on the feedforward link from the BS to the UEs, where information about the uplink precoders as well as the uplink pilot sequences is periodically transmitted. In general, a quantization scheme can be defined as follows. Let $\boldsymbol{x} \in \mathbb{C}^d$ denote the quantized source, i.e. either an uplink precoder or a pilot sequence. In both cases, the knowledge of $\boldsymbol{x}$ or $\boldsymbol{x}e^{i\varphi}$ for any phase φ are equivalent. Moreover, we consider that the power $\|\boldsymbol{x}\|^2$ is either fed back separately from $\boldsymbol{x}$, or allocated at the UE. Thus, the information on the phase or amplitude does not matter, i.e. we want to quantize a source $\boldsymbol{x}$ that lies in the Grassmannian set of lines in $\mathbb{C}^d$ denoted by $\mathcal{G}$. Let B denote the total number of bits used for the quantization of $\boldsymbol{x}$. The distortion of a quantizer is measured as the expected error between the source vector $\boldsymbol{x}$ and its quantized version.

Some of the authors recently introduced [29] a vector quantization method for Grassmannian variables, called cube-split, which enables high-resolution (large B) quantization at low computational complexity for large vector dimensions (large d), while remaining efficient from a distortion point of view if the source $\boldsymbol{x}$ is uniformly drawn on the Grassmannian. The rationale of the proposed quantization approach is the following: the considered Grassmannian space (homogeneous to a sphere) is first split into cells looking like bent hypercubes (hence the cube-split name); then a bent lattice is defined on each cell through a mapping chosen such that the resulting codewords are approximately uniformly distributed on the Grassmannian. This approach performs favorably versus scalar quantization of the components of $\boldsymbol{x}$—the classical choice in complexity-constrained situations—while exhibiting the same complexity. In our approach, cube-split quantization is used on the feed-forward links to transmit the uplink precoders from the BS to the UEs.

B. Multi-User Precoding

As shown on Figure 2, in our approach the BS centrally computes the set of scheduled UL users $\mathcal{U} \subseteq \{1, \dots K\}$ and the corresponding UL precoders $\{\mathbf{W}_u\}_{u \in \mathcal{U}}$. Here we restrict

the precoders to follow a multiple-beamforming strategy with uniform power allocation, i.e.,

$$\mathbf{W}_u = \sqrt{P_u/S_u}\left[\mathbf{v}_{u,\mathcal{S}_u(1)}, \ldots, \mathbf{v}_{u,\mathcal{S}_u(S_u)}\right], \quad u \in \mathcal{U} \quad (17)$$

where P_u is the transmit power constraint of the u-th UE, $\{\mathbf{v}_{u,s}\}_{s=1}^{r_u}$ are the right eigenvectors of $\mathbf{H}_u$, and $\mathcal{S}_u(s)$ denotes the s-th element of the set of scheduled eigenmodes for the u-th UE, $\mathcal{S}_u \subseteq \{1, \ldots, r_u\}$, with cardinality $S_u = |\mathcal{S}_u| \leq r_u$.

The set of scheduled UL users $\mathcal{U}$ and the corresponding sets of scheduled eigenmodes $\{\mathcal{S}_u\}_{u\in\mathcal{U}}$ are selected to maximize the sum-rate under linear equalization and independent per-stream decoding.

We adopt for our setup a greedy approach following [30] which first schedules the substream experiencing the best channel conditions and, then, as long as the sum-rate increases, schedules the substream with best signal-to-interference- and-noise ratio (SINR), given the set of already scheduled users.

C. Simulation Channel model

To obtain realistic channels, we adopt an implementation of the Third Generation Partnership Project (3GPP) channel model, from specifications in [31]. The 3GPP channel model is a well accepted simulation platform for 3G and 4G applications, especially its latest iterations. This channel model is linked to a system-level simulator matching the Long Term Evolution (LTE) specifications and parameters [28]. The main simulation parameters are described in Table I.

Considering our application, the channels that we simulate should have some of the expected characteristics of Massive MIMO [10], [32]. For example, since the multipath components of the channels generated according to [31] are bounded in the angular domain, they will have a rank-deficient covariances as per Lemma 2 in [9]. This is especially true of line-of-sight channels where secondary rays have negligible power. However, we have found that the 3GPP channel model is lacking some core characteristics to truthfully capture the covariance dynamics of real networks. From [23] and subsequent work in the channel measurement community, we know that covariance matrices are not entirely stationary over time and frequency for a given scenario. In the 3GPP model, due to the way the channels are computed and since the impinging angles of each multipath component remain constant during user movement, the covariance also stays constant over time and over frequency. For the current work, our evaluation is restricted to a relatively small portion of the time-frequency grid. We can thus expect that the covariance is indeed constant in both intervals, and that our covariance tracking will approach the true value.

Rigorously evaluating these algorithm without resorting to prototyping would require a consistent—and hopefully tractable—non-stationary simulation model, with covariance evolving in space and in frequency [10]. Up to the authors knowledge, such a model has yet to be validated and thus provides an interesting research area for the near future.

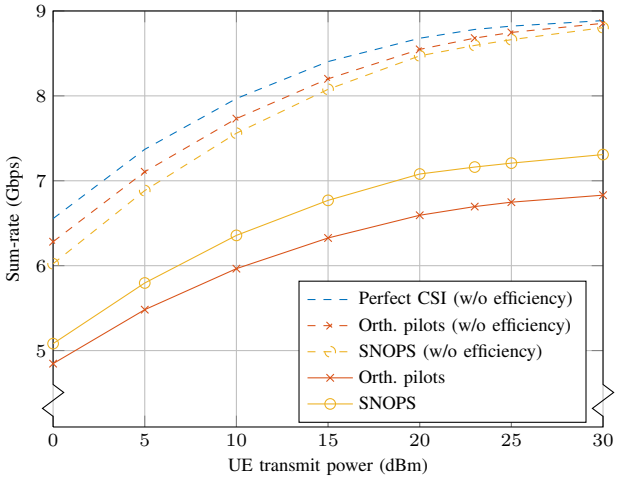

Fig. 3. Comparison between full pilots and shorter, non-orthogonal pilots from [19].

D. Simulation results

We evaluate the performance of the proposed architecture by considering the achieved uplink throughput, evaluated from the SINR computed based on the quantized uplink precoders, true channel, and LMMSE multi-user equalization with estimated CSI at the BS. Out total bandwidth is 200 MHz, over which we apply uplink precoders on 1 MHz subbands. In a classical LTE systems, this corresponds to around 60 subcarriers. On all the figures in the follow-up discussion, we consider here the greedy precoding method, channel model and the simulation parameters from Section IV.

On Fig. 3, we compare the achieved sum-rate performance of our system, with and without efficiency (16) factored in. The chosen decoding bandwidth is set at 0.5 MHz, i.e. half the precoding bandwidth. Without any efficiency consideration, the performance of both orthogonal pilots and non-orthogonal pilots in this case are very close to the genie-aided perfect CSI information over the considered SNR range. As can be expected, using shorter pilots lead to a small

TABLE I
CHANNEL PARAMETERS CONSIDERED FOR THE SIMULATION.

Parameters	Value
# of UEs	12
Parameters set (from [31])	Urban Macro
BS height	25 m
Maximum distance	100 m
User speed	3 km/h
Pathloss model	from [31]
Line-of-sight probability	0.6
# of BS antennas	64 (Rectangular)
# of UE antennas	8 (Rectangular)
Shadowing	Yes, from [31]
Frame structure	LTE compliant [28]
Bandwidth	200 MHz (10×20 MHz)
Duplexing	TDD
Uplink:Downlink frames	2:2 (LTE config. 1 [28])

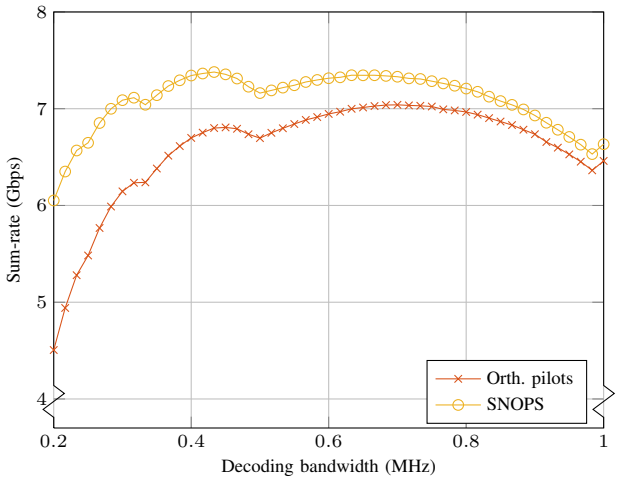

Fig. 4. Comparison of the sum-rate achieved with full pilots and shorter, non-orthogonal pilots from [19]. We consider different granularities of decoding bandwidth, with sounding efficiency factored in.

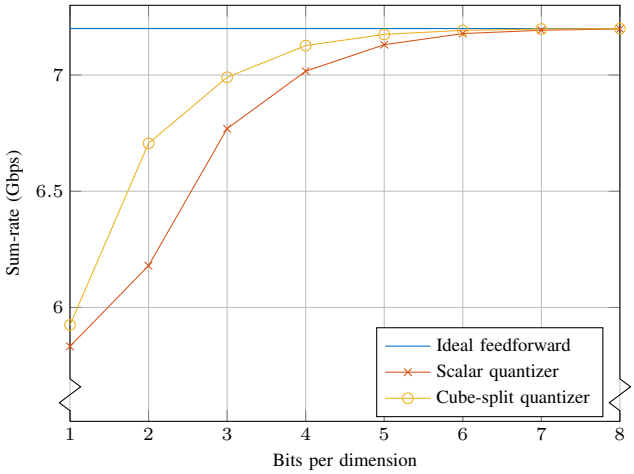

Fig. 5. Performance of the feedforward quantizers, comparing the cube-split quantizer [29] with a baseline scalar quantizer with a similar number of bits per dimension.

decrease in performance—once again in an absolute sense. However, factoring in efficiency gives a strong advantage to the shorter pilots. Using our simulation parameters, the orthogonal pilots had length $N \cdot L = 96$ time-frequency resources, whereas short pilots used an average of 70 symbols for similar performance. The efficiency of orthogonal pilots is then $\eta_{\text{orth}} = 1 - 96/420 = 0.77$, whereas the efficiency of the short pilots is $\eta_{\text{SNOPS}} = 0.83$.

This gain in performance is obviously linked to the choice of decoding bandwidth, as discussed in the introduction of this section. Consider Fig. 4, where we compare the throughput of the system with a increasingly larger decoding bandwidth starting from a single resource block of 200 KHz. At this extreme end, orthogonal pilots have a very low efficiency of around 0.5 and the gains of using shorter pilots is most pronounced. However, as expected, the advantages of using shorter pilots decrease when the decoding bandwidth increases. Note that the variation of performance over the decoding bandwidth is also linked to the channel correlation over frequency; similar effects can be observed e.g. in [23].

Finally, we analyze the effect of feedforward quantization on the overall UL precoding scheme on Fig. 5. The interest here lies in the number of bits per real dimension of the quantized precoding vectors that are needed in order to perform close to the ideal case. As in the case of shorter pilots, a lower number of bits per real dimension improves the efficiency of the system in the feedforward channel. We can readily see that for a low number of bits, quantization noise has a high impact on the overall performance of the system. In the range of interest, between 2 to 4 bits per real dimension, the cube-split quantizer operates closer to the ideal feedforward case.

V. Multi-Cell Architecture

The extension of the proposed design to the multi-cell case can be done in a straightforward manner. Although we do not present a full analysis of this case, there are some indications that the proposed design can be scaled to large multi-cell networks; the main argument is that a distributed design applicable to multiple cells and capable of handling inter-cell interference can be devised under fairly mild assumptions. In more detail, let us assume that i) significant inter-cell interference is only coming from a small number of neighbor cells, regardless of the total network size; ii) channel estimation is performed synchronously across all cells; and iii) information about the pilot allocation to the users is shared among these neighboring BS. Note that we do not require the pilot allocation to be done jointly or collaboratively; we merely assume that the pilot sequences are locally shared.

The channel and covariance estimation approaches previously described in Sections II-B and III respectively can be applied in a distributed fashion to each cell, while including the interfering users from neighboring cells in the covariance tracking and channel estimation process. In other words, thanks only to assumptions ii) and iii) above, each BS can keep track of both the statistics and instantaneous channels estimates of the significant inter-cell interferers.

This CSI can then be used by the BS for interference-aware multi-user equalization and/or decoding during the transmission of uplink data, and for intra and inter-cell interference nulling during downlink transmission. The design principle advocated here is that once inter-cell CSI is available, inter-cell interference can be mitigated with little or no cooperation; no further exchange of user data, nor of statistical or instantaneous CSI is required. Therefore, the proposed approach can scale to networks of arbitrary size, without the overhead in design complexity or processing typically associated with joint data processing or centralized precoder computation.

VI. Conclusion

We considered CSI acquisition strategies suitable for Massive MIMO TDD systems based on the tracking of channel spatial statistics to reduce the pilot overhead. We evaluated

the associated practical application and performance limits, accounting for covariance estimation and realistic channel models, by proposing a complete design addressing all these issues. Our results show that when the cost of CSI acquisition is properly accounted for, the tracking of statistical CSI together with the use of non-orthogonal pilots can enable to achieve a higher effective spectral efficiency than what can be achieved with orthogonal pilot sequences.

REFERENCES

[1] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[2] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, feb 2014.

[3] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, feb 2013.

[4] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, apr 2003.

[5] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[6] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[7] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving "Massive MIMO" spectral efficiency with a not-so-large number of antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3226–3239, Sep. 2012.

[8] J. Hoydis, C. Hoek, T. Wild, and S. ten Brink, "Channel measurements for large antenna arrays," in *Proc. IEEE International Symposium on Wireless Communication Systems (ISWCS)*, 2012.

[9] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE Journ. Sel. Areas in Comm.*, vol. 31, no. 2, pp. 264–273, feb 2013.

[10] P. Ferrand, M. Amara, S. Valentin, and M. Guillaud, "Trends and challenges in wireless channel modeling for evolving radio access," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 93–99, jul 2016.

[11] J. H. Kotecha and A. M. Sayeed, "Transmit signal design for optimal estimation of correlated MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 546–557, Feb. 2004.

[12] E. Björnson and B. Ottersten, "A framework for training-based estimation in arbitrarily correlated Rician MIMO channels with Rician disturbance," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1807–1820, mar 2010.

[13] N. Shariati, J. Wang, and M. Bengtsson, "Robust training sequence design for correlated MIMO channel estimation," *IEEE Transactions on Signal Processing,*, vol. 62, no. 1, pp. 107–120, jan 2014.

[14] A. Assalini, E. DallAnese, and S. Pupolin, "Linear MMSE MIMO channel estimation with imperfect channel covariance information," in *Proc. International Conference on Communications (ICC)*, 2009.

[15] A. Adhikary, J. Nam, J. Ahn, and G. Caire, "Joint spatial division and multiplexing: The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.

[16] S. E. Hajri, M. Assaad, and G. Caire, "Scheduling in massive MIMO: User clustering and pilot assignment," in *Proc. Allerton Conference on Communication, Control, and Computing*, sep 2016.

[17] I. Atzeni, J. Arnau, and M. Debbah, "Fractional pilot reuse in massive MIMO systems," in *Proc. International Conference on Communications (ICC)*, 2015.

[18] E. Björnson, J. Hoydis, and L. Sanguinetti, "Pilot contamination is not a fundamental asymptotic limitation in massive MIMO," 2017, submitted to IEEE ICC. [Online]. Available: https://arxiv.org/abs/1611.09152

[19] B. Tomasi and M. Guillaud, "Pilot length optimization for spatially correlated multi-user MIMO channel estimation," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, 2015.

[20] M. Guillaud, D. T. M. Slock, and R. Knopp, "A practical method for wireless channel reciprocity exploitation through relative calibration," in *Proc. Eighth International Symposium on Signal Processing and Its Applications (ISSPA '05)*, Sydney, Australia, 2005.

[21] B. Tomasi, A. Decurninge, and M. Guillaud, "SNOPS: Short non-orthogonal pilot sequences for downlink channel state estimation in FDD massive MIMO," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, Washington DC, USA, 2016.

[22] H. Q. Ngo and E. G. Larsson, "No downlink pilots are needed in TDD massive MIMO," *IEEE Transactions on Transactions on Wireless Communications*, 2017, accepted.

[23] I. Viering, H. Hofstetter, and W. Utschick, "Validity of spatial covariance matrices over time and frequency," in *Proc. IEEE Global Telecommunications Conf. GLOBECOM '02*, vol. 1, Nov. 2002, pp. 851–855.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[25] E. Björnson, L. Sanguinetti, and M. Debbah. (2016) Massive MIMO with imperfect channel covariance information. [Online]. Available: https://arxiv.org/pdf/1612.04128v2.pdf

[26] Y. Chi, Y. Eldar, and R. Calderbank, "Petrels: Parallel subspace estimation and tracking by recursive least squares from partial observations," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5947–5959, dec 2013.

[27] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 303–318, jan 2017.

[28] C. Cox, *An introduction to LTE*. John Wiley & Sons, 2014.

[29] A. Decurninge and M. Guillaud, "Cube-split: Structured quantizers on the Grassmannian of lines," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2017. [Online]. Available: https://arxiv.org/abs/1610.08257

[30] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[31] 3GPP Technical Specification Group Radio Access Network, "Study on 3D channel model for LTE (release 12)," Tech. Rep. TR 36.873, 2015.

[32] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO performance evaluation based on measured propagation data," *IEEE Transactions on Wireless Communications*, vol. 14, no. 7, pp. 3899–3911, jul 2015.