Information Theory and Wireless Channel Modeling

Mérouane Debbah[1]

Alcatel-Lucent Chair on Flexible Radio, SUPELEC, 3 rue Joliot-Curie 91192 GIF SUR YVETTE CEDEX, France
merouane.debbah@supelec.fr

1 Introduction

The problem of modelling channels is crucial for the efficient design of wireless systems [1, 2, 3]. The wireless channel suffers from constructive/destructive interference signaling [4, 5]. This yields a randomized channel with certain statistics to be discovered. Recently ([6, 7]), the need to increase spectral efficiency has motivated the use of multiple antennas at both the transmitter and the receiver side. Hence, in the case of i.i.d Gaussian entries of the MIMO link and perfect channel knowledge at the receiver, it has been proved [8] that the ergodic capacity increase is $\min(n_r,n_t)$ bits per second per hertz for every 3dB increase (n_r is the number of receiving antennas and n_t is the number of transmitting antennas) at high Signal to Noise Ratio (SNR)[1]. However, for realistic[2] channel models, results are still unknown and may seriously put into doubt the MIMO hype. As a matter of fact, the actual design of efficient codes is tributary of the channel model available: the transmitter has to know in what environment the transmission occurs in order to provide the codes with the adequate properties: as a typical example, in Rayleigh fading channels, when coding is performed, the Hamming distance (also known as the number of distinct components of the multi-dimensional constellation) plays a central role whereas maximizing the Euclidean distance is the commonly approved design criteria for Gaussian channels (see Giraud and Belfiore [9] or Boutros and Viterbo [10]).

As a consequence, channel modelling is the key in better understanding the limits of transmissions in wireless and noisy environments. In particular, questions of the form: "what is the highest transmission rate on a propagation environment where we only know the mean of each path, the variance of each path and the directions of arrival?" are crucially important. It will justify the use (or not) of MIMO technologies for a given state of knowledge.

Let us first introduce the modelling constraints. We assume that the transmission takes place between a mobile transmitter and receiver. The transmitter has n_t antennas and the receiver has n_r antennas. Moreover, we assume that the input transmitted signal goes through a time variant linear filter channel. Finally, we assume that the interfering noise is additive white Gaussian.

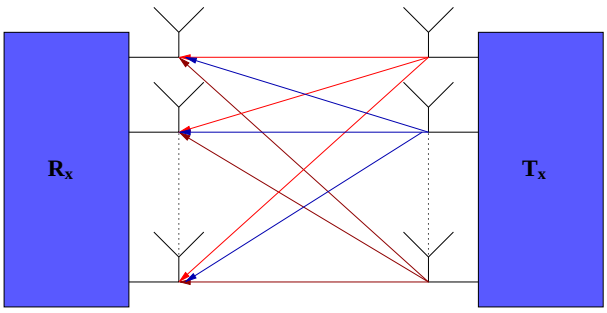

Fig. 1. MIMO channel representation.

[1] In the single antenna Additive White Gaussian Noise (AWGN) channel, 1 bit per second per hertz can be achieved with every 3dB increase at high SNR.

[2] By realistic, we mean models representing our state of knowledge of reality which might be different from reality.

The transmitted signal and received signal are related as:

$$\mathbf{y}(t) = \sqrt{\frac{\rho}{n_t}} \int \mathbf{H}_{n_r \times n_t}(\tau, t)\mathbf{x}(t-\tau)d\tau + \mathbf{n}(t) \tag{1}$$

with

$$\mathbf{H}_{n_r \times n_t}(\tau, t) = \int \mathbf{H}_{n_r \times n_t}(f, t)e^{j2\pi f \tau} df \tag{2}$$

ρ is the received SNR (total transmit power per symbol versus total spectral density of the noise), t, f and τ denote respectively time, frequency and delay, $\mathbf{y}(t)$ is the $n_r \times 1$ received vector, $\mathbf{x}(t)$ is the $n_t \times 1$ transmit vector, $\mathbf{n}(t)$ is an $n_r \times 1$ additive standardized white Gaussian noise vector.

In the rest of the paper, we will only be interested in the frequency domain modelling (knowing that the impulse response matrix can be accessed through an inverse Fourier transform according to relation 2). We would like to provide some theoretical grounds to model the frequency response matrix $\mathbf{H}(f, t)$ based on a given state of knowledge. In other words, knowing only certain things related to the channel (Directions of Arrival (DoA), Directions of Departure (DoD), bandwidth, center frequency, number of transmitting and receiving antennas, number of chairs in the room...), how to attribute a joint probability distribution to the entries $h_{ij}(f, t)$ of the matrix:

$$\mathbf{H}_{n_r \times n_t}(f, t) = \begin{pmatrix} h_{11}(f,t) & \ldots\ldots & h_{1n_t}(f,t) \\ \vdots & \ldots\ldots & \vdots \\ \vdots & \ldots\ldots & \vdots \\ h_{n_r 1}(f,t) & \ldots\ldots & h_{n_r n_t}(f,t) \end{pmatrix} \tag{3}$$

This question can be answered in light of the Bayesian probability theory and the principle of maximum entropy. Bayesian probability theory has led to a profound theoretical understanding of various scientific areas [11, 12, 13, 14, 15, 16, 17, 18] and has shown the potential of entropy as a measure of our degree of knowledge when encountering a new problem. The principle of maximum entropy[3] is at present the clearest theoretical justification in conducting scientific inference: we do not need a model, entropy maximization creates a model for us out of the information available. Choosing the distribution with greatest entropy avoids the arbitrary introduction or assumption of information that is not available[4]. Bayesian probability theory improves on maximum entropy by expressing some prior knowledge on the model and estimating the parameters of the model.

As we will emphasize all along this paper, channel modelling is not a science representing reality but only our knowledge of reality as thoroughly stated by Jaynes in [20]. It answers in particular the following question: based on a given state of knowledge (usually brought by raw data or prior information), what is the best model one can make? This is, of course, a vague question since there is no strict definition of what is meant by best. But what do we mean then by best? In this contribution, our aim is to derive a model which is adequate with our state of knowledge. We need a measure of uncertainty which expresses the constraints of our knowledge and the desire to leave the unknown parameters to lie in an unconstrained space. To this end, many possibilities are offered to us to express our uncertainty. However, we need an information measure which is consistent (complying to certain common sense desiderata, see [21] to express these desiderata and for the derivation of entropy) and easy to manipulate: we need a general principle for translating information into probability assignment. Entropy is the measure of information that fulfills this criteria. Hence, already in 1980, Shore et al. [21] proved that the principle of maximum entropy is the correct method of inference when given new information in terms of expected values. They proved that maximizing entropy is correct in the following sense: maximizing any function but entropy will lead to inconsistencies unless that function and entropy have the same maximum[5]. The consistency argument is at the heart of scientific inference and can be expressed through the following axiom[6]:

[3] The principle of maximum entropy was first proposed by Jaynes [12, 13] as a general inference procedure although it was first used in Physics.

[4] Keynes named it judiciously the principle of indifference [19] to express our indifference in attributing prior values when no information is available.

[5] Thus, aiming for consistency, we can maximize entropy without loss of generality.

[6] The consistency property is only one of the required properties for any good calculus of plausibility statement. In fact, R.T Cox in 1946 derived three requirements known as Cox's Theorem[22]:

Lemma 1. *If the prior information* $\mathbf{I_1}$ *on which is based the channel model* $\mathbf{H_1}$ *can be equated to the prior information* $\mathbf{I_2}$ *of the channel model* $\mathbf{H_2}$ *then both models should be assigned the same probability distribution* $P(\mathbf{H}) = P(\mathbf{H_1}) = P(\mathbf{H_2})$.

Any other procedure would be inconsistent in the sense that, by changing indices 1 and 2, we could then generate a new problem in which our state of knowledge is the same but in which we are assigning different probabilities. More precisely, Shore et al. [21] formalize the maximum entropy approach based on four consistency axioms stated as follows[7]:

- Uniqueness: If one solves the same problem twice the same way then the same answer should result both times.
- Invariance: If one solves the same problem in two different coordinate systems then the same answer should result both times.
- System independence: It should not matter whether one accounts for independent information about independent systems separately in terms of different densities or together in terms of a joint density.
- Subset independence: It should not matter whether one treats an independent subset of system states in terms of a separate conditional density or in terms of the full system density.

These axioms are based on the fundamental principle that if a problem can be solved in more than one way, the results should be consistent. Given this statement in mind, the rules of probability theory should lead every person to the same unique solution, provided each person bases his model on the same information.[8]

Moreover, the success over the years of the maximum entropy approach (see Boltzmann's kinetic gas law, [23] for the estimate of a single stationary sinusoidal frequency, [14] for estimating the spectrum density of a stochastic process subject to autocorrelation constraints, [24] for estimating parameters in the context of image reconstruction and restoration problems, [25] for applying the maximum entropy principle on solar proton event peak fluxes in order to determine the least biased distribution) has shown that this information tool is the right way so far to express our uncertainty.

Let us give an example in the context of spectral estimation of the powerful feature of the maximum entropy approach which has inspired this paper. Suppose a stochastic process x_i for which $p+1$ autocorrelation values are known i.e $\mathbb{E}(x_i x_{i+k}) = \tau_k, k = 0, ..., p$ for all i. What is the consistent model one can make of the stochastic process based only on that state of knowledge, in other words the model which makes the least assumption on the structure of the signal? The maximum entropy approach creates for us a model and shows that, based on the previous information, the stochastic process is a p^{th} auto-regressive (AR) order model process of the form [14]:

$$x_i = -\sum_{k=1}^{p} a_k x_{i-k} + b_i$$

where the b_i are i.i.d zero mean Gaussian distributed with variance σ^2 and $a_1, a_2, .., a_p$ are chosen to satisfy the autocorrelation constraints (through Yule-Walker equations).

- Divisibility and comparability: the plausibility of of a statement is a real number between 0 (for false) and 1 (for true) and is dependent on information we have related to the statement.
- Common sense: Plausibilities should vary with the assessment of plausibilities in the model.
- Consistency: If the plausibility of a statement can be derived in two ways, the two results should be equal.

[7] In all the rest of the document, the consistency argument will be referred to as Axiom 1.

[8] It is noteworthy to say that if a prior distribution Q of the estimated distribution P is available in addition to the expected values constraints, then the principle of minimum cross-entropy (which generalizes maximum entropy) should be applied. The principle states that, of the distribution P that satisfy the constraints, one should choose the one which minimizes the functional:

$$D(P,Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

Minimizing cross-entropy is equivalent to maximizing entropy when the prior Q is a uniform distribution. Intuitively, cross-entropy measures the amount of information necessary to change the prior Q into the posterior P. If measured data is available, Q can be estimated. However, one can only obtain a numerical form for P in this case (which is not always useful for optimization purposes). Moreover, this is not a easy task for multidimensional vectors such as vec($\mathbf{H}$). As a consequence, we will always assume a uniform prior and use therefore the principle of maximum entropy.

In this contribution, we would like to provide guidelines for creating models from an information theoretic point of view and therefore make extensive use of the principle of maximum entropy together with the principle of consistency.

2 Some Considerations

2.1 Channel Modelling Methodology

In this contribution, we provide a methodology (already successfully used in Bayesian spectrum analysis [23, 17]) for inferring on channel models. The goal of the modelling methodology is twofold:

- to define a set of rules, called hereafter *consistency axioms*, where only our state of knowledge needs to be defined.
- to use a measure of uncertainty, called hereafter *entropy*, in order to avoid the arbitrary introduction or assumption of information that is not available.

In other words, if two published papers make the same assumptions in the abstract (concrete buildings in Oslo where one avenue...), then both papers should provide the same channel model.

To achieve this goal, in all this document, the following procedure will be applied: every time we have some information on the environment (*and not make assumptions on the model*!), we will ask a question based on that the information and provide a model taking into account that information and nothing more! The resulting model and its compliance with later test measurements will justify whether the information used for modelling was adequate to characterize the environment in sufficient details. Hence, when asked the question, "what is the consistent model one can make knowing the directions of arrival, the number of scatterers, the fact that each path has zero mean and a given variance?" we will suppose that the information provided by this question is unquestionable and true i.e the propagation environment depends on fixed steering vectors, each path has effectively zero mean and a given variance. We will suppose that effectively, when waves propagate, they bounce onto scatterers and that the receiving antenna sees these ending scatterers through steering directions. Once we assume this information to be true, we will construct the model based on Bayesian tools.[9].

To explain this point of view, the author recalls an experiment made by his teacher during a tutorial explanation on the duality behavior of light: photon or wave. The teacher took two students of the class, called here

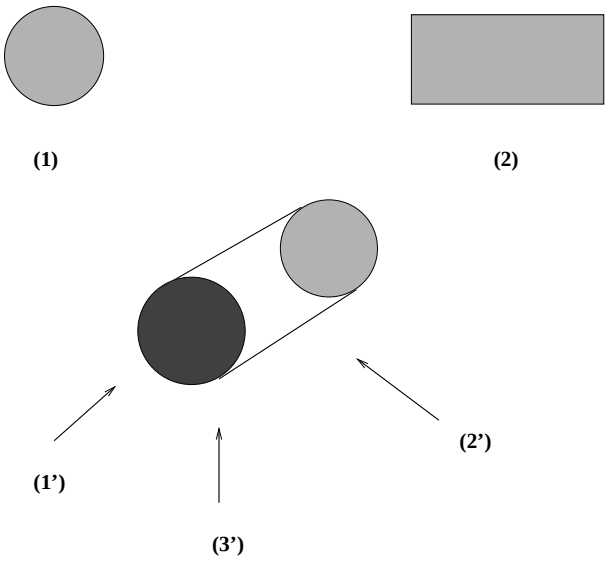

Fig. 2. Duality wave-corpuscule?

A and B for simplicity sake. To student A, he showed view (1') (see Figure 2) of a cylinder and to student

[9] Note that in Bayesian inference, all probabilities are conditional on some hypothesis space (which is assumed to be true).

B, he showed view (2') of the same cylinder. For A, the cylinder was a circle and for B, the cylinder was a rectangle. Who was wrong? Well, nobody. Based on the state of knowledge (1'), representing the cylinder as a circle is the best one can do. Any other representation of the cylinder would have been made on unjustified assumptions (the same applies to view (2')). Unless we have another state of knowledge (view (3')), the true nature of the object will not be found.

Our channel modelling will not pretend to seek reality but only to represent view (1') or view (2') in the most accurate way (i.e if view (1') is available then our approach should lead into representing the cylinder as a circle and not as a triangle for example). If the model fails to comply with measurements, we will not put into doubt the model but conclude that the information we had at hand to create the model was insufficient. We will take into account the failure as a new source of information and refine/change our question in order to derive a new model based on the principle of maximum entropy which complies with the measurements. This procedure will be routinely applied until the right question (and therefore the right answer) is found. When performing scientific inference, every question asked, whether right or wrong, is important. Mistakes are eagerly welcomed as they lead the path to better understand the propagation environment. Note that the approach devised here is not new and has already been used by Jaynes [20] and Jeffrey [26]. We give hereafter a summary of the modelling approach:

1. **Question selection**: the modeler asks a question based on the information available.
2. **Construct the model**: the modeler uses the principle of maximum entropy (with the constraints of the question asked) to construct the model M_i.
3. **Test**: (When complexity is not an issue) The modeler computes the a posteriori probability of the model and ranks the model.
4. **Return to 1.**: The outcome of the test is some "new information" evidence to keep/refine/change the question asked. Based on this information, the modeler can therefore make a new model selection.

This algorithm is iterated as many times as possible until better ranking is obtained. However, we have to alert the reader on one main point: the convergence of the previous algorithm is not at all proven. Does this mean that we have to reject the approach? we should not because our aim is to better understand the environment and by successive tests, we will discard some solutions and keep others.

We provide hereafter a brief historical example to highlight the methodology. In the context of spectrum estimation, the Schuster periodogram (also referred in the literature as the discrete Fourier transform power spectrum) is commonly used for the estimation of hidden frequencies in the data. The Schuster periodogram is defined as:

$$F(\omega) = \frac{1}{N} \mid \sum_{k=1}^{N} s_k e^{-j\omega t_k} \mid^2$$

s_k is the data of length N to be analyzed. In order to find the hidden frequencies in the data, the general procedure is to maximize $F(\omega)$ with respect to ω . But as in our case, one has to understand why/when to use the Schuster periodogram for frequency estimation. The Schuster periodogram answers a specific question based on a specific assumption (see the work of Bretthorst [17]). In fact, it answers the following question: "what is the optimal frequency estimator for a data set which contains a **single stationary sinusoidal frequency** in the presence of Gaussian white noise?" From the standpoint of Bayesian probability, the discrete Fourier Transform power spectrum answers a specific question about single (and not two or three....) stationary sinusoidal frequency estimation. Given this state of knowledge, the periodogram will consider everything in the data that cannot be fit to a single sinusoid to be noise and will therefore, if other frequencies are present, misestimate them. However, if the periodogram does not succeed in estimating multiple frequencies, the periodogram is not to blame but only the question asked! One has to devise a new model (a model maybe based on a two stationary sinusoidal frequencies?). This new model selection will lead to a new frequency estimator in order to take into account the structure of what was considered to be noise. This routine is repeated and each time, the models can be ranked to determine the right number of frequencies.

2.2 Information and Complexity

In the introduction, we have recalled the work of Shore et al. [21] which shows that maximizing entropy leads to consistent solutions. However, incorporating information in the entropy criteria which is not given in terms of expected values is not an easy task. In particular, how does one incorporate information on the fact that the room has four walls and two chairs? In this case, we will not maximize entropy based only on

the information we have (expected values and number of chairs and walls): we will maximize entropy based on the expected values and a structured form of the channel matrix (which is more than the information we have since the chairs and walls are not constraint equations in the entropy criteria). This ad-hoc procedure will be used because it is extremely difficult to incorporate knowledge on physical considerations (number of chairs, type of room...) in the entropy criteria. Each time this ad-hoc procedure is used, we will verify that although we maximize entropy under a structured constraint, we remain consistent. Multiple iterations of this procedure will refine the structured form of the channel until the modeler obtains a consistent structured models that maximizes entropy.

A question the reader could ask is whether we should take into account all the information provided, in other words, what information is useful? We should of course consider all the available information but there is a compromise to be made in terms of model complexity. Each information added will not have the same effect on the channel model and might as well more complicate the model for nothing rather than bring useful insight on the behavior of the propagation environment. To assume further information by putting some additional structure would not lead to incorrect predictions: however, if the predictions achieved with or without the details are equivalent, then this means that the details may exist but are irrelevant for the understanding of our model[10]. As a typical example, when conducting iterative decoding analysis [27], Gaussian models of priors are often sufficient to represent our information. Inferring on other moments and deriving the true probabilities will only complicate the results and not yield a better understanding.

3 Gaussian i.i.d Channel Model

3.1 Finite Energy Case

In this section, we give a precise justification on why and when the Gaussian i.i.d model should be used. We recall the general model:

$$\mathbf{y} = \sqrt{\frac{\rho}{n_t}}\mathbf{H}\mathbf{x} + \mathbf{n}$$

Imagine now that the modeler is in a situation where it has no measurements and no knowledge where the transmission took place. The only thing the modeler knows is that the channel carries some energy E, in other words, $\frac{1}{n_r n_t}\mathbb{E}\left(\sum_{i=1}^{n_r}\sum_{j=1}^{n_t} \mid h_{ij} \mid^2\right) = E$. Knowing only this information, the modeler is faced with the following question: what is the consistent model one can make knowing only the energy E (but not the correlation even though it may exist) ? In other words, based on the fact that:

$$\int d\mathbf{H} \sum_{i=1}^{n_r}\sum_{j=1}^{n_t} \mid h_{ij} \mid^2 P(\mathbf{H}) = n_t n_r E \qquad \text{(Finite energy)} \tag{4}$$

$$\int dP(\mathbf{H}) = 1 \qquad \text{(P(}\mathbf{H}\text{) is a probability distribution)} \tag{5}$$

What distribution $P(\mathbf{H})$[11] should the modeler assign to the channel? The modeler would like to derive the most general model complying with those constraints, in other words the one which maximizes our uncertainty while being certain of the energy. This statement can simply be expressed if one tries to maximize the following expression using Lagrange multipliers with respect to P:

$$L(P) = -\int d\mathbf{H} P(\mathbf{H}) \log P(\mathbf{H}) + \gamma \sum_{i=1}^{n_r}\sum_{j=1}^{n_t}[E - \int d\mathbf{H} \mid h_{ij} \mid^2 P(\mathbf{H})] + \beta\left[1 - \int d\mathbf{H} P(\mathbf{H})\right]$$

[10] Limiting one's information is a general procedure that can be applied to many other fields. As a matter of fact, the principle "one can know less but understand more" seems the only reasonable way to still conduct research considering the huge amount of papers published each year.

[11] It is important to note that we are concerned with $P(\mathbf{H} \mid I)$ where I represents the general background knowledge (here the variance) used to formulate the problem. However, for simplicity sake, $P(\mathbf{H} \mid I)$ will be denoted $P(\mathbf{H})$.

If we derive $L(P)$ with respect to P, we get:

$$\frac{dL(P)}{dP} = -1 - \log P(\mathbf{H}) - \gamma \sum_{i=1}^{n_r} \sum_{j=1}^{n_t} \mid h_{ij} \mid^2 - \beta = 0$$

then this yields:

$$\begin{aligned} P(\mathbf{H}) &= e^{-(\beta+\gamma \sum_{i=1}^{n_r} \sum_{j=1}^{n_t} |h_{ij}|^2 + 1)} \\ &= e^{-(\beta+1)} \prod_{i=1}^{n_r} \prod_{j=1}^{n_t} \exp -(\gamma \mid h_{ij} \mid^2) \\ &= \prod_{i=1}^{n_r} \prod_{j=1}^{n_t} P(h_{ij}) \end{aligned}$$

with

$$P(h_{ij}) = e^{-(\gamma |h_{ij}|^2 + \frac{\beta+1}{n_r n_t})}.$$

One of the most important conclusions of the maximum entropy principle is that while we have only assumed the variance, these assumptions imply independent entries since the joint probability distribution $P(\mathbf{H})$ simplifies into products of $P(h_{ij})$. Therefore, based on the previous state of knowledge, the only maximizer of the entropy is the i.i.d one. This does not mean that we have supposed independence in the model. In the generalized $L(P)$ expression, there is no constraint on the independence. Another surprising result is that the distribution achieved is Gaussian. Once again, gaussianity is not an assumption but a consequence of the fact that the channel has finite energy. The previous distribution is the least informative probability density function that is consistent with the previous state of knowledge. When only the variance of the channel paths are known (but not the frequency bandwidth, nor knowledge of how waves propagate, nor the fact that scatterers exist...) then the only consistent model one can make is the Gaussian i.i.d model. In order to fully derive $P(\mathbf{H})$, we need to calculate the coefficients β and γ. The coefficients are solutions of the following constraint equations:

$$\int d\mathbf{H} \sum_{i=1}^{n_r} \sum_{j=1}^{n_t} \mid h_{ij} \mid^2 P(\mathbf{H}) = n_t n_r E$$

$$\int d\mathbf{H} P(\mathbf{H}) = 1$$

Solving the previous equations yields the following probability distribution:

$$P(\mathbf{H}) = \frac{1}{(\pi E)^{n_r n_t}} \exp\{-\sum_{i=1}^{n_r} \sum_{j=1}^{n_t} \frac{\mid h_{ij} \mid^2}{E}\}$$

Of course, if one has any additional knowledge, then this information should be integrated in the $L(P)$ criteria and would lead to a different result.
As a typical example, suppose that the modeler knows that the frequency paths have different variances such as $\mathbb{E}(\mid h_{ij} \mid^2) = E_{ij}$. Using the same methodology, it can be shown that :

$$P(\mathbf{H}) = \prod_{i=1}^{n_r} \prod_{j=1}^{n_t} P(h_{ij})$$

with $P(h_{ij}) = \frac{1}{\pi E_{ij}} e^{-\frac{|h_{ij}|^2}{E_{ij}}}$. The principle of maximum entropy still attributes independent Gaussian entries to the channel matrix but with different variances.
Suppose now that the modeler knows that the path h_{pk} has a mean equal to $\mathbb{E}(h_{pk}) = m_{pk}$ and variance $\mathbb{E}(\mid h_{pk} - m_{pk} \mid^2) = E_{pk}$, all the other paths having different variances (but nothing is said about the mean). Using as before the same methodology, we show that:

$$P(\mathbf{H}) = \prod_{i=1}^{n_r} \prod_{j=1}^{n_t} P(h_{ij})$$

with for all $\{i,j,(i,j)\neq(p,k)\}$ $P(h_{ij})=\frac{1}{\pi E_{ij}}e^{-\frac{|h_{ij}|^2}{E_{ij}}}$ and $P(h_{pk})=\frac{1}{\pi E_{pk}}e^{-\frac{|h_{pk}-m_{pk}|^2}{E_{pk}}}$. Once again, different but still independent Gaussian distributions are attributed to the MIMO channel matrix.

The previous examples can be extended and applied whenever a modeler has some new source of information **in terms of expected values** on the propagation environment[12]. In the general case, if N constraints are given on the expected values of certain functions $\int g_i(\mathbf{H})P(\mathbf{H})d\mathbf{H}=\alpha_i$ for $i=1...N$, then the principle of maximum entropy attributes the following distribution [28]:

$$P(\mathbf{H})=e^{-(1+\lambda+\sum_{i=1}^{N}\lambda_i g_i(\mathbf{H}))}$$

where the values of λ and λ_i (for $i=1..N$) can be obtained by solving the constraint equations.

Although these conclusions are widely known in the Bayesian community, the author is surprised that many MIMO channel papers begin with: "let us assume a $n_r\times n_t$ matrix with Gaussian i.i.d entries...". No assumptions on the model should be made. Only the state of knowledge should be clearly stated at the beginning of each paper and the conclusion of the maximum entropy approach can be straightforwardly used.[13]

As a matter of fact, the Gaussian i.i.d model should not be "thrown" away but be extensively used whenever our information on the propagation conditions is scarce (we don't know in what environment we are transmitting our signal i.e the frequency, the bandwidth, WLAN scenario, we do not know what performance measure we target...)[14].

3.2 Finite Energy unknown

We will consider a case similar to the previous section where the modeler is in a situation where it has no measurements and no knowledge where the transmission took place. The modeler does know that the channel carries some energy E but is not aware of its value.

In the case where the modeler knows the value of E, we have shown that:

$$P(\mathbf{H}\mid E)=\frac{1}{(\pi E)^{n_r n_t}}\exp\{-\sum_{i=1}^{n_r}\sum_{j=1}^{n_t}\frac{\mid h_{ij}\mid^2}{E}\}$$

In general, when E is unknown, the probability distribution is derived according to:

$$\begin{aligned}P(\mathbf{H})&=\int P(\mathbf{H},E)dE\\&=\int P(\mathbf{H}\mid E)P(E)dE\end{aligned}$$

and is consistent with the case where E is known i.e $P(E)=\delta(E-E_0)$:

$$P(\mathbf{H})=\frac{1}{(\pi E_0)^{n_r n_t}}\exp\{-\sum_{i=1}^{n_r}\sum_{j=1}^{n_t}\frac{\mid h_{ij}\mid^2}{E_0}\}$$

In the case were the energy E is unknown, one has to determine $P(E)$. E is a positive variance parameter and the channel can not carry more energy than what is transmitted (i.e $E\leq E_{\max}$) . This is merely the sole knowledge the modeler has about E on which the modeler has to derive a prior distribution[15].

[12] The case where information is not given in terms of expected values is treated afterwards.

[13] "Normality is not an assumption of physical fact at all. It is a valid description of our state of information", Jaynes.

[14] In "The Role of Entropy in Wave Propagation" [29], Franceschetti et al. show that the probability laws that describe electromagnetic magnetic waves are simply maximum entropy distributions with appropriate moment constraints. They suggest that in the case of dense lattices, where the inter-obstacle hitting distance is small compared to the distance traveled, the relevant metric is non-Euclidean whereas in sparse lattices, the relevant metric becomes Euclidean as propagation is not constrained along the axis directions.

[15] Jeffrey [26] already in 1939 proposed a way to handle this issue based on invariance properties and consistency axioms. He suggested that a proper way to express incomplete ignorance of a continuous variable known to be positive is to assign uniform prior probability to its logarithm, in other words: $P(E)\propto\frac{1}{E}$. However, the distribution is improper and one can not therefore marginalize with this distribution.

In this case, using maximum entropy arguments, one can derive $P(E)$:

$$P(E) = \frac{1}{E_{\max}} \qquad 0 \leq E \leq E_{\max}$$

As a consequence,

$$P(\mathbf{H}) = \int_0^{E_{\max}} \frac{1}{(\pi E)^{n_r n_t}} \exp\{-\sum_{i=1}^{n_r}\sum_{j=1}^{n_t} \frac{\mid h_{ij} \mid^2}{E}\} dE$$

With the change of variables $u = \frac{1}{E}$, we obtain:

$$P(\mathbf{H}) = \frac{1}{E_{\max}\pi^{n_r n_t}} \int_{\frac{1}{E_{\max}}}^{\infty} u^{n_r n_t - 2} e^{-\sum_{i=1}^{n_r}\sum_{j=1}^{n_t}|h_{ij}|^2 u} du$$

Note that the distribution is invariant by unitary transformations, is not Gaussian and moreover the entries are not independent when the modeler has no knowledge on the amount of energy carried by the channel. This point is critical and shows the effect of the lack of information on the exact energy[16].

In the case $n_t = 1$ and $n_r = 2$, we obtain:

$$P(\mathbf{H}) = \frac{1}{E_{\max}\pi^2 \sum_{i=1}^{2} \mid h_{i1} \mid^2} e^{-\frac{\sum_{i=1}^{2}|h_{i1}|^2}{E_{\max}}}$$

3.3 Correlation matrix unknown

Suppose now that the modeler knows that correlation exists between the entries of the channel matrix $\mathbf{H}$ but is not aware of the value of the correlation matrix $\mathbf{Q} = \mathbb{E}(\mathrm{vec}(\mathbf{H})\mathrm{vec}(\mathbf{H})^H)$. What consistent distribution should the modeler attribute to the channel based only on that knowledge?

To answer this question, suppose that the correlation matrix $\mathbf{Q} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^H$ is known ($\mathbf{V} = [\mathbf{v}_1, ...\mathbf{v}_{n_r n_t}]$ is a $n_r n_t \times n_r n_t$ unitary matrix whereas $\mathbf{\Lambda}$ is a $n_r n_t \times n_r n_t$ diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_{n_r n_t})$ with $\lambda_i \geq 0$ for $1 \leq i \leq n_r n_t$).

Using the maximum entropy principle, one can easily show that:

$$P(\mathbf{H} \mid \mathbf{V}, \mathbf{\Lambda}) = \frac{1}{\prod_{i=1}^{n_r n_t} \pi\lambda_i} \exp\{\sum_{i=1}^{n_r n_t} \frac{\mid \mathbf{v_i}^H \mathrm{vec}(\mathbf{H}) \mid^2}{\lambda_i}\}$$

The channel distribution can be obtained:

$$\begin{aligned} P(\mathbf{H}) &= \int P(\mathbf{H}, \mathbf{V}, \mathbf{\Lambda}) d\mathbf{V} d\mathbf{\Lambda} \\ &= \int P(\mathbf{H} \mid \mathbf{V}, \mathbf{\Lambda}) P(\mathbf{V}, \mathbf{\Lambda}) d\mathbf{V} d\mathbf{\Lambda} \end{aligned}$$

If the correlation matrix is perfectly known, then $P(\mathbf{V}, \mathbf{\Lambda}) = \delta(\mathbf{V} - \mathbf{V}^0)\delta(\mathbf{\Lambda} - \mathbf{\Lambda}^0)$ and

$$P(\mathbf{H}) = \frac{1}{\prod_{i=1}^{n_r n_t} \pi{\lambda^0}_i} \exp\{\sum_{i=1}^{n_r n_t} \frac{\mid \mathbf{v^0_i}^H \mathrm{vec}(\mathbf{H}) \mid^2}{{\lambda^0}_i}\}$$

In the case were the correlation matrix $\mathbf{Q}$ is unknown, one has to determine $P(\mathbf{V}, \mathbf{\Lambda}) = P(\mathbf{\Lambda} \mid \mathbf{V})P(\mathbf{V})$. This is the problem of constructing an ignorance prior corresponding to ignorance of both scale (up to some constraints proper to our problem) and rotation. The a priori distribution can be derived as well as the joint probability distribution using tools from statistical physics. Due to limited space, the result is not provided but can be found in the recent work of the author [30].

[16] In general, closed form solutions of the distributions do not exist. In this case, a powerful tool for approximate Bayesian inference that uses Markov Chain Monte Carlo to compute marginal posterior distributions of interest can be used through WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml.)

4 Knowledge of the directions of arrival, departure, delay, bandwidth, power: frequency selective channel model with time variance

4.1 Knowledge of the directions of arrival or departure

The modeler[17] is interested in modelling the channel over time scales over which the locations of scatterers do not not change significantly relative to the transmitter or receiver. This is equivalent to considering time scales over which the channel statistics do not change significantly. However, the channel realizations do vary over such time scales. Imagine that the modeler is in a situation where it knows the energy carried by the channel (nothing is known about the mean)[18]. Moreover, the modeller knows from electromagnetic theory that when a wave propagates from a scatterer to the receiving antennas, the signal can be written in an exponential form

$$\mathbf{s}(t,\mathbf{d}) = \mathbf{s_0}\, e^{j(\mathbf{k}^T\mathbf{d}-2\pi ft)} \tag{6}$$

which is the plane wave solution of the Maxwell equations in free non-dispersive space for wave vector $\mathbf{k} \in \Re^{2\times 1}$ and location vector $\mathbf{d} \in \Re^{2\times 1}$. The reader must note that other solutions to the Maxwell equations exist and therefore the modeler is making an important restriction. The direction of the vector $\mathbf{s_0}$ gives us knowledge on the polarization of the wave while the direction of the wave vector $\mathbf{k}$ gives us knowledge on the direction of propagation. The phase of the signal results in $\phi = \mathbf{k}^T\mathbf{d}$. The modeler considers for simplicity sake that the scatterers and the antennas lie in the same plane. The modeler makes use of the knowledge that the steering vector is known up to a multiplicative complex constant that is the same for all antennas.

Although correlation might exist between the scatterers, the modeler is not aware of such a thing. Based on this state of knowledge, the modeler wants to derive a model which takes into account all the previous constraints while leaving as many degrees of freedom as possible to the other parameters (since the modeler does not want to introduce unjustified information). In other words, based on the fact that:

$$\mathbf{H} = \frac{1}{\sqrt{s_r}} \begin{pmatrix} e^{j\phi_{1,1}} & \dots & e^{j\phi_{1,s_r}} \\ \vdots & \ddots & \vdots \\ e^{j\phi_{n_r,1}} & \dots & e^{j\phi_{n_r,s_r}} \end{pmatrix} \mathbf{\Theta}_{s_r\times n_t}$$

what distribution should the modeler attribute to $\mathbf{\Theta}_{s_r\times n_t}$? $\mathbf{H}$ is equal to $\frac{1}{\sqrt{s_r}}\mathbf{\Phi\Theta}$, $\phi_{i,j} = \mathbf{k}.\mathbf{r}_{i,j}$ and $\mathbf{r}_{i,j}$ is the distance between the receiving antenna i and receiving scatterer j and $\mathbf{\Phi}$is a $n_r \times s_r$ matrix (s_r is the number of scatterers) which represents the directions of arrival from randomly positioned scatterers to the receiving antennas. $\mathbf{\Theta}_{s_r\times n_t}$ is an $s_r \times n_t$ matrix which represents the scattering environment between the transmitting antennas and the scatterers (see Figure 3).

The consistency argument (see Proposition 1) states that if the DoA (Directions of Arrival) are unknown then $\mathbf{H} = \frac{1}{\sqrt{s_r}}\mathbf{\Phi}_{n_r\times s_r}\mathbf{\Theta}_{s_r\times n_t}$ should be assigned an i.i.d Gaussian distribution since the modeler is in the same state of knowledge as before where it only knew the variance.

Based on the previous remarks, let us now derive the distribution of $\mathbf{\Theta}_{s_r\times n_t}$. The probability distribution $P(\mathbf{H})$ is given by:

$$P(\mathbf{H}) = \int P(\mathbf{\Phi\Theta} \mid \mathbf{\Phi}, s_r) P(\mathbf{\Phi} \mid s_r) P(s_r) ds_r d\mathbf{\Phi}$$

- When $\mathbf{\Phi}$ and s_r are known, then $P(\mathbf{\Phi} \mid s_r) = \delta(\mathbf{\Phi} - \mathbf{\Phi}^0)$ and $P(s_r) = \delta(s_r - {s_r}^0)$. Therefore $P(\mathbf{H}) = P(\mathbf{\Phi}^0\mathbf{\Theta})$.
- When $\mathbf{\Phi}$ and s_r are unknown: the probability distribution of the frequency path h_{ij} is:

$$P(h_{ij}) = \int P\left(h_{ij} \mid \mathbf{\Phi}, s_r\right) P(\mathbf{\Phi} \mid s_r) P(s_r) d\mathbf{\Phi} ds_r \tag{7}$$

In the case when $P(\mathbf{\Phi} \mid s_r)$ and $P(s_r)$ are unknown, the consistency argument states that:

[17] We treat in this section thoroughly the directions of arrival model and show how the directions of departure model can be easily obtained from the latter case.

[18] The case where the paths have different non-zero means can be treated the same way.

- The $\mathbf{\Theta}_{s_r \times n_t}$ matrix is such as each h_{ij} is zero mean Gaussian.
- The $\mathbf{\Theta}_{s_r \times n_t}$ matrix is such as $\mathbb{E}(h_{ij} {h_{mn}}^*) = \delta_{im}\delta_{jn}$ (since h_{ij} is Gaussian, decorrelation is equivalent to independence).

In this case, the following result holds:

Proposition 1. $\mathbf{\Theta}_{s_r \times n_t}$ *i.i.d. zero mean Gaussian with unit variance is solution of the consistency argument and maximizes entropy.*

Proof: Since $\mathbf{\Phi}$ is unknown, the principle of maximum entropy attributes independent uniformly distributed angles to each entry ϕ_{ij}:

$$P(\phi_{ij}) = \frac{1}{2\pi}\mathbf{1}_{[0,2\pi]}.$$

Let us show that $\mathbf{\Theta}_{s_r \times n_t}$ i.i.d zero mean with variance 1 is solution of the consistency argument.

Since $h_{ij} = \frac{1}{\sqrt{s_r}}\sum_{k=1}^{s_r}\theta_{kj}e^{j\phi_{ik}}$ then $P(h_{ij} \mid \mathbf{\Phi}, s_r) = N(0, \frac{1}{s_r}\sum_{k=1}^{s_r} \mid e^{j\phi_{ik}} \mid^2 = 1) = \frac{1}{\sqrt{2\pi}}e^{-\frac{|h_{ij}|^2}{2}}$ and therefore h_{ij} is zero mean Gaussian since:

$$\begin{aligned}
P(h_{ij}) &= \int P\left(h_{ij} \mid \mathbf{\Phi}, s_r\right) P(\mathbf{\Phi} \mid s_r) P(s_r) d\mathbf{\Phi} ds_r \\
&= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{|h_{ij}|^2}{2}} P(\mathbf{\Phi} \mid s_r) P(s_r) d\mathbf{\Phi} ds_r \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{|h_{ij}|^2}{2}} \int P(\mathbf{\Phi} \mid s_r) P(s_r) d\mathbf{\Phi} ds_r \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{|h_{ij}|^2}{2}}
\end{aligned}$$

Moreover, we have:

$$\begin{aligned}
\mathbb{E}(h_{ij} h_{mn}^*) &= \mathbb{E}_{\mathbf{\Theta},\mathbf{\Phi}}\left(\frac{1}{\sqrt{s_r}}\sum_{k=1}^{s_r}\theta_{kj}e^{j\phi_{ik}} \frac{1}{\sqrt{s_r}}\sum_{l=1}^{s_r}\theta_{ln}^* e^{-j\phi_{ml}}\right) \\
&= \frac{1}{s_r}\sum_{k=1}^{s_r}\sum_{l=1}^{s_r}\mathbb{E}_{\mathbf{\Theta}}(\theta_{kj}\theta_{ln}^*)\mathbb{E}_{\mathbf{\Phi}}(e^{j\phi_{ik}-j\phi_{ml}}) \\
&= \frac{1}{s_r}\sum_{k=1}^{s_r}\sum_{l=1}^{s_r}\delta_{kl}\delta_{jn}\mathbb{E}_{\mathbf{\Phi}}(e^{j\phi_{ik}-j\phi_{ml}}) \\
&= \delta_{jn}\frac{1}{s_r}\sum_{k=1}^{s_r}\mathbb{E}_{\mathbf{\Phi}}(e^{j\phi_{ik}-j\phi_{mk}}) \\
&= \delta_{jn}\delta_{im}
\end{aligned}$$

which proves that $\mathbf{H}$ is i.i.d Gaussian for unknown angles.

One interesting point of the maximum entropy approach is that while we have not assumed uncorrelated scattering, the above methodology will automatically assign a model with uncorrelated scatterers in order to have as many degrees of freedom as possible. But this does not mean that correlation is not taken into account. The model in fact leaves free degrees for correlation to exist or not. The maximum entropy approach is appealing in the sense that if correlated scattering is given as a prior knowledge, then it can be immediately integrated in the channel modelling approach (as a constraint on the covariance matrix for example). Note also that in this model, the entries of $\mathbf{H}$ are correlated for general DoA's.

Suppose now that the modeler assumes that the different steering vectors have different amplitudes $\sqrt{P_i^r}$. What distribution should the modeler attribute to the matrix $\mathit{\Theta}_{s_r \times n_t}$ in the following representation:

$$\mathbf{H} = \frac{1}{\sqrt{s_r}}\begin{pmatrix} e^{j\phi_{1,1}} & \dots & e^{j\phi_{1,s_r}} \\ \vdots & \ddots & \vdots \\ e^{j\phi_{n_r,1}} & \dots & e^{j\phi_{n_r,s_r}} \end{pmatrix}\begin{pmatrix} \sqrt{P_1^r} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sqrt{P_{s_r}^r} \end{pmatrix}\mathbf{\Theta}_{s_r \times n_t}?$$

Proposition 2. $\mathbf{\Theta}_{s_r \times n_t}$ *i.i.d Gaussian with variance 1 is solution of the consistency argument and maximizes entropy*

Proof: We will not go into the details as the proof is a particular case of the proof of Proposition 3.

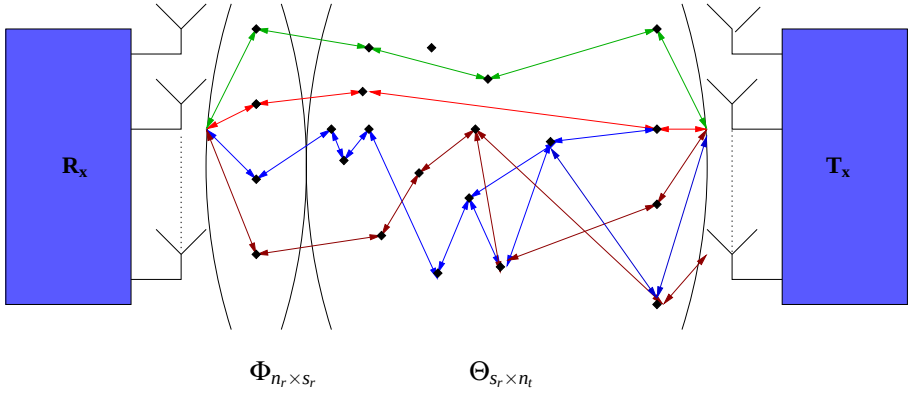

Fig. 3. Directions of arrival based model.

4.2 Knowledge of the Directions of Arrival and Departure

The modeler is now interested in deriving a consistent double directional model i.e taking into account simultaneously the directions of arrival and the directions of departure. The motivation of such an approach lies in the fact that when a single bounce on a scatterer occurs, the direction of arrival and departure are deterministically related by Descartes laws and therefore the distribution of the channel matrix depends on the joint DoA-DoD spectrum. The modeler assumes as a state of knowledge the directions of departure from the transmitting antennas to the set of transmitting scatterers $(1...s_t)$. The modeler also assumes as a state of knowledge the directions of arrival from the set of receiving scatterers $(1...s_r)$ to the receiving antennas. The modeler also has some knowledge that the steering directions have different powers. However, the modeler has no knowledge of what happens in between. The set $(1...s_t)$ and $(1...s_r)$ may be equal, $(1...s_t)$ may be included in $(1...s_r)$ or there may be no relation between the two. The modeler also knows that the channel carries some energy. Based on this state of knowledge, what is the consistent model the modeler can make of $\mathbf{H}$

$$\mathbf{H} = \frac{1}{\sqrt{s_r s_t}} \begin{pmatrix} e^{j\phi_{1,1}} & \dots & e^{j\phi_{1,s_r}} \\ \vdots & \ddots & \vdots \\ e^{j\phi_{n_r,1}} & \dots & e^{j\phi_{n_r,s_r}} \end{pmatrix} \begin{pmatrix} \sqrt{P_1^r} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sqrt{P_{s_r}^r} \end{pmatrix}$$
$$\mathbf{\Theta}_{s_r \times s_t} \begin{pmatrix} \sqrt{P_1^t} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sqrt{P_{s_t}^t} \end{pmatrix} \begin{pmatrix} e^{j\psi_{1,1}} & \dots & e^{j\psi_{1,n_t}} \\ \vdots & \ddots & \vdots \\ e^{j\psi_{s_t,1}} & \dots & e^{j\psi_{s_t,n_t}} \end{pmatrix} ?$$

In other words, how to model $\mathbf{\Theta}_{s_r \times s_t}$? As previously stated, the modeler must comply with the following constraints:

- The channel has a certain energy.
- Consistency argument: If the DoD and DoA are unknown then $\frac{1}{\sqrt{s_r s_t}} \mathbf{\Phi}_{n_r \times s_r} \mathbf{P^r}^{\frac{1}{2}} \mathbf{\Theta}_{s_r \times s_t} \mathbf{P^t}^{\frac{1}{2}} \mathbf{\Psi}_{s_t \times n_t}$ should be assigned an i.i.d zero mean Gaussian distribution.

  Let us now determine the distribution of $\mathbf{\Theta}_{s_r \times s_t}$. The probability distribution of $P(\mathbf{H})$ is given by:

$$P(\mathbf{H}) = \int P(\mathbf{\Phi P^r}^{\frac{1}{2}} \mathbf{\Theta P^t}^{\frac{1}{2}} \mathbf{\Psi} \mid \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{P^r}, \mathbf{P^t}, s_r, s_t) P(\mathbf{\Psi}, \mathbf{\Phi} \mid s_r, s_t) P(\mathbf{P^r}, \mathbf{P^t} \mid s_t, s_r) P(s_t, s_r) ds_r ds_t d\mathbf{P^r} d\mathbf{P^t} d\mathbf{\Psi} d\mathbf{\Phi}$$

- When $\mathbf{\Psi}, \mathbf{\Phi}, s_r, s_t, \mathbf{P^r}, \mathbf{P^t}$ are known: $P(\mathbf{\Phi\Psi} \mid s_r, s_t) = \delta(\mathbf{\Phi} - \mathbf{\Phi}^0)\delta(\mathbf{\Psi} - \mathbf{\Psi}^0)$, $P(s_t, s_r) = \delta(s_r - {s^0}_r)\delta(s_t - {s^0}_t)$,$P(\mathbf{P^r}, \mathbf{P^t} \mid s_r, s_t) = \delta(\mathbf{P^r} - \mathbf{P^{0}}^r)\delta(\mathbf{P^t} - \mathbf{P^{0}}^t)$ and

$$P(\mathbf{H}) = P(\mathbf{\Phi}^0 \mathbf{P^{0}}^{r\,\frac{1}{2}} \mathbf{\Theta P^{0}}^{t\,\frac{1}{2}} \mathbf{\Psi}^0)$$

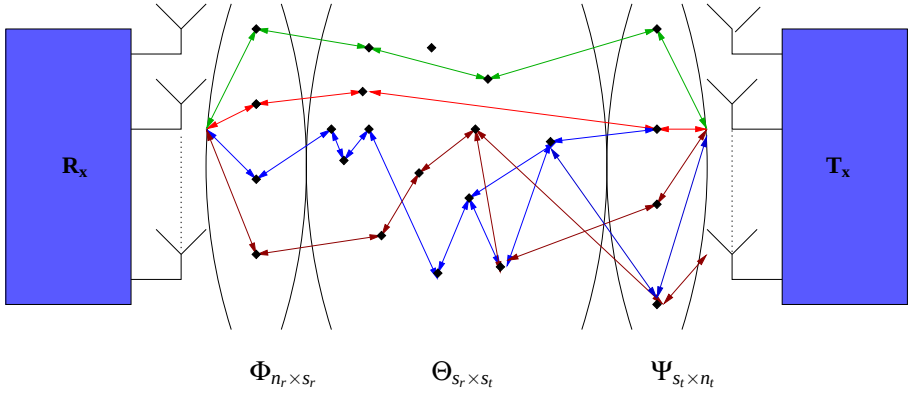

Fig. 4. Double directional based model.

- Suppose now that $\mathbf{\Psi}, \mathbf{\Phi}, s_r, s_t$ are unknown, then each entry h_{ij} of $\mathbf{H}$ must have an i.i.d zero mean Gaussian distribution. In this case, the following result holds:

  Proposition 3. $\mathbf{\Theta}_{s_r \times s_t}$ *i.i.d zero mean Gaussian with variance* 1 *is solution of the consistency argument and maximizes entropy.*

  Proof: Let us show that $\mathbf{\Theta}_{s_r \times s_t}$ i.i.d zero mean Gaussian with variance 1 is solution of the consistency argument and maximizes entropy. Since $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are unknown, the principle of maximum entropy attributes i.i.d uniform distributed angles over 2π to the entries ϕ_{ij} and ψ_{ij}. In this case, if one chooses $\theta_{p,k}$ to be i.i.d zero mean Gaussian with variance 1 and knowing that $h_{ij} = \frac{1}{\sqrt{s_t s_r}} \sum_{k=1}^{s_t} \sum_{p=1}^{s_r} \theta_{pk} \sqrt{P_k{}^t} \sqrt{P_p{}^r} e^{j\psi_{kj}} e^{j\phi_{ip}}$, then: $P\left(h_{ij} \mid \mathbf{\Psi}, \mathbf{\Phi}, s_r, s_t\right) = N(0, \frac{1}{s_t s_r} \sum_{p=1}^{s_r} \sum_{k=1}^{s_t} \mid \sqrt{P_p{}^r} e^{j\phi_{ip}} \sqrt{P_k{}^t} e^{j\psi_{kj}} \mid^2 = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{|h_{ij}|^2}{2}}$ (since $\frac{1}{s_r} \sum_{k=1}^{s_r} P_k{}^t = 1$ and $\frac{1}{s_t} \sum_{p=1}^{s_t} P_p{}^t = 1$ (due to power normalization as we assume the energy known). Therefore

$$
\begin{aligned}
P(h_{ij}) &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{|h_{ij}|^2}{2}} P(\mathbf{\Phi}, \mathbf{\Psi} \mid s_t, s_r) P(\mathbf{P^r}, \mathbf{P^t} \mid s_t, s_r) P(s_t, s_r) d\mathbf{\Phi} d\mathbf{\Psi} \\
&\quad d\mathbf{P^r} d\mathbf{P^t} ds_t ds_r \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{|h_{ij}|^2}{2}} \int P(\mathbf{\Phi}, \mathbf{\Psi} \mid s_t, s_r) P(\mathbf{P^r}, \mathbf{P^t} \mid s_t, s_r) P(s_t, s_r) d\mathbf{\Phi} d\mathbf{\Psi} d\mathbf{P^r} d\mathbf{P^t} ds_t ds_r \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{|h_{ij}|^2}{2}}
\end{aligned}
$$

  Moreover, we have :

$$
\begin{aligned}
\mathbb{E}_{\mathbf{\Phi},\mathbf{\Psi},\mathbf{\Theta}}(h_{ij} h_{mn}^*) &= \frac{1}{s_t s_r} \sum_{k=1}^{s_t} \sum_{p=1}^{s_r} \sum_{r=1}^{s_t} \sum_{l=1}^{s_r} \mathbb{E}_{\mathbf{\Theta}}(\theta_{pk} \theta_{lr}^*) \mathbb{E}_{\Psi}(e^{-j\psi_{rn} + j\psi_{kj}}) \mathbb{E}_{\Phi}(e^{-j\phi_{ml} + j\phi_{ip}}) \\
&\quad \sqrt{P_k{}^t} \sqrt{P_r{}^t} \sqrt{P_p{}^r} \sqrt{P_l{}^r} \\
&= \frac{1}{s_t s_r} \sum_{k=1}^{s_t} \sum_{p=1}^{s_r} \sum_{r=1}^{s_t} \sum_{l=1}^{s_r} \delta_{pl} \delta_{kr} \mathbb{E}_{\Psi}(e^{-j\psi_{rn} + j\psi_{kj}}) \mathbb{E}_{\Phi}(e^{-j\phi_{ml} + j\phi_{ip}}) \\
&\quad \sqrt{P_k{}^t} \sqrt{P_r{}^t} \sqrt{P_p{}^r} \sqrt{P_l{}^r} \\
&= \frac{1}{s_t s_r} \sum_{k=1}^{s_t} \sum_{p=1}^{s_r} \mathbb{E}_{\Psi}(e^{-j\psi_{kn} + j\psi_{kj}}) \mathbb{E}_{\Phi}(e^{-j\phi_{mp} + j\phi_{ip}}) P_k{}^t P_p{}^r \\
&= \delta_{im} \delta_{jn} \frac{1}{s_t s_r} \sum_{k=1}^{s_t} \sum_{p=1}^{s_r} P_k{}^t P_p{}^r \\
&= \delta_{im} \delta_{jn}
\end{aligned}
$$

which proves that $\mathbf{\Theta}_{s_r \times s_t}$ is solution of the consistency argument. Once again, instead of saying that this model represents a rich scattering environment, it should be more correct to say that the model makes allowance for every case that could be present to happen since we have imposed no constraints besides the energy.

4.3 Considering more features

The modeler wants to derive a consistent model taking into account the direction of arrivals and respective power profile, directions of departure and respective power profile, delay, Doppler effect. As a starting point, the modeler assumes that the position of the transmitter and receiver changes in time. However, the scattering environment (the buildings, trees,...) does not change and stays in the same position during the transmission. Let $\mathbf{v_t}$ and $\mathbf{v_r}$ be respectively the vector speed of the transmitter and the receiver with respect to a terrestrial reference (see Figure 5). Let $\mathbf{s}_{ij}^t$ be the signal between the transmitting antenna i and the first scatterer j. Assuming that the signal can be written in an exponential form (plane wave solution of the Maxwell equations) then:

$$\begin{aligned}\mathbf{s}_{ij}^t(t) &= \mathbf{s_0} e^{j(\mathbf{k}_{ij}^t(\mathbf{v_t}t+\mathbf{d_{ij}})+2\pi f_c t)} \\ &= \mathbf{s_0} e^{j2\pi(\frac{f_c \mathbf{u}_{ij}^t \mathbf{v_t}}{c}t+f_c t)} e^{j\psi_{ij}}\end{aligned}$$

Here, f_c is the carrier frequency, $\mathbf{d}_{ij}$ is the initial vector distance between antenna i and scatterer j ($\psi_{ij} = \mathbf{k}_{ij}^t.\mathbf{d}_{ij}$ is the scalar product between vector $\mathbf{k}_{ij}^t$ and vector $\mathbf{d}_{ij}$), $\mathbf{k}_{ij}^t$ is such as $\mathbf{k}_{ij}^t = \frac{2\pi}{\lambda}\mathbf{u}_{ij}^t = \frac{2\pi f_c}{c}\mathbf{u}_{ij}^t$. The quantity $\frac{1}{2\pi}\mathbf{k}_{ij}^t \mathbf{v_t}$ represents the Doppler effect.

In the same vein, if we define $\mathbf{s}_{ij}^r(t)$ as the signal between the receiving antenna j and the scatterer i, then:

$$s_{ij}^r(t) = \mathbf{s_0} e^{j(2\pi(\frac{f_c \mathbf{v_r} \mathbf{u}_{ij}^r}{c}t+f_c t))} e^{j\phi_{ij}}$$

In all the following, the modeler supposes as a state of knowledge the following parameters:

- speed $\mathbf{v_r}$.
- speed $\mathbf{v_t}$.
- the angle of departure from the transmitting antenna to the scatterers ψ_{ij} and power P_j^t.
- the angle of arrival from the scatterers to the receiving antenna ϕ_{ij} and power P_j^r.

The modeler has however no knowledge of what happens in between except the fact that a signal going from a steering vector of departure j to a steering vector of arrival i has a certain delay τ_{ij} due to possible single bounce or multiple bounces on different objects. The modeler also knows that objects do not move between the two sets of scatterers. The $s_r \times s_t$ delay matrix linking each DoA and DoD has the following structure:

$$\mathbf{D}_{s_r \times s_t}(f) = \begin{pmatrix} e^{-j2\pi f \tau_{1,1}} & \dots & e^{-j2\pi f \tau_{1,s_t}} \\ \vdots & \ddots & \vdots \\ e^{-j2\pi f \tau_{s_r,1}} & \dots & e^{-j2\pi f \tau_{s_r,s_t}} \end{pmatrix}$$

The modeler also supposes as a given state of knowledge the fact that each path h_{ij} of matrix $\mathbf{H}$ has a certain power. Based on this state of knowledge, the modeler wants to model the $s_r \times s_t$ matrix $\mathbf{\Theta}_{s_r \times s_t}$ in the following representation:

$$\begin{aligned}\mathbf{H}(f,t) = \frac{1}{\sqrt{s_r s_t}} &\begin{pmatrix} e^{j(\phi_{1,1}+2\pi\frac{f\mathbf{u}_{11}^r\mathbf{v_r}}{c}t)} & \dots & e^{j(\phi_{1,s}+2\pi\frac{f\mathbf{u}_{1s}^r\mathbf{v_r}}{c}t)} \\ \vdots & \ddots & \vdots \\ e^{j(\phi_{r,1}+2\pi\frac{f\mathbf{u_{r1}}^r\mathbf{v_r}}{c}t)} & \dots & e^{j(\phi_{r,s}+2\pi\frac{f\mathbf{u}_{rs}^r\mathbf{v_r}}{c}t)} \end{pmatrix} \begin{pmatrix} \sqrt{P_1^r} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sqrt{P_{s_r}^r} \end{pmatrix} \\ &\mathbf{\Theta}_{s_r \times s_t} \bigodot \mathbf{D}_{s_r \times s_t}(f) \\ &\begin{pmatrix} \sqrt{P_1^t} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sqrt{P_{s_t}^t} \end{pmatrix} \begin{pmatrix} e^{j(\psi_{1,1}+2\pi\frac{f\mathbf{u}_{11}^t\mathbf{v_t}}{c}t)} & \dots & e^{j(\psi_{1,n_t}+2\pi\frac{f\mathbf{u}_{1n_t}^t\mathbf{v_t}}{c}t)} \\ \vdots & \ddots & \vdots \\ e^{j(\psi_{s_1,1}+2\pi\frac{f\mathbf{u}_{s_1 1}^t\mathbf{v_t}}{c}t)} & \dots & e^{j(\psi_{s_1,n_t}+2\pi\frac{f\mathbf{u}_{s_1 n_t}^t\mathbf{v_t}}{c}t)} \end{pmatrix}\end{aligned}$$

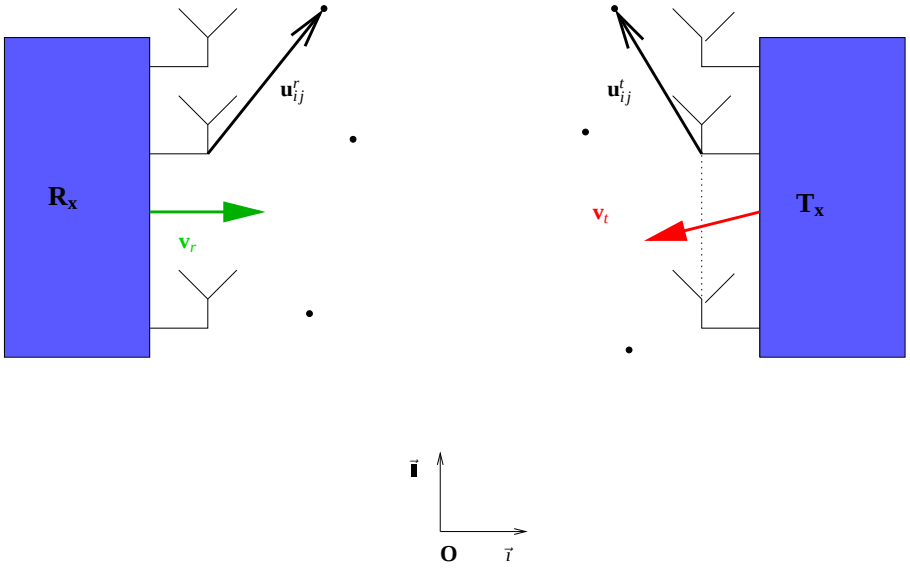

Fig. 5. Moving antennas.

$\odot$ represents the Hadamard product defined as $c_{ij} = a_{ij}b_{ij}$ for a product matrix $\mathbf{C} = \mathbf{A} \odot \mathbf{B}$. As previously stated, one has to comply with the following constraints:

- Each entry of $\mathbf{H}(f,t)$ has a certain energy.
- Consistency argument: if the DoA, DoD, powers, the delays, the Doppler effects are unknown then matrix $\mathbf{H}$ should be assigned an i.i.d Gaussian distribution.

Proposition 4. $\mathbf{\Theta}_{s_r \times s_t}$ *i.i.d zero mean Gaussian with variance 1 is solution of the consistency argument and maximizes entropy.*[19]

Proof: We will not go into the details but only provide the guidelines of the proof. First, remark that if $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are unknown, then the principle of maximum entropy attributes i.i.d uniform distribution to the angles ϕ_{ij} and ψ_{ij}. But what probability distribution should the modeler attribute to the delays and the Doppler effects when no information is available?

- **Delays**: The modeler knows that there is, due to measurements performed in the area, a maximum possible delay for the information to go from the transmitter to the receiver $\tau_{\max}$. The principle of maximum entropy attributes therefore a uniform distribution to all the delays τ_{ij} such as $P(\tau_{ij}) = \frac{1}{\tau_{\max}}$ with $\tau_{ij} \in [0, \tau_{\max}]$
- **Doppler effect**: The modeler knows that the speed of the transmitter and receiver can not exceed a certain limit v_{limit} (in the least favorable case, v_{limit} would be equal to the speed of light) but if the transmission occurs in a city, the usual car speed limit can be taken as an upper bound. In this case, the speed v_t and v_r have also a uniform distribution such as $P(v_t) = P(v_r) = \frac{1}{v_{\text{limit}}}$. Moreover, if $\mathbf{v}_t = v_t \cos(\alpha_t)\boldsymbol{\imath} + v_t \sin(\alpha_t)\boldsymbol{\jmath}$, $\mathbf{v}_r = v_r \cos(\alpha_r)\boldsymbol{\imath} + v_r \sin(\alpha_r)\boldsymbol{\jmath}$, $\mathbf{u^t}_{ij} = \cos({\beta^t}_{ij})\boldsymbol{\imath} + \sin({\beta^t}_{ij})\boldsymbol{\jmath}$ and $\mathbf{u^r}_{ij} = \cos({\beta^r}_{ij})\boldsymbol{\imath} + \sin({\beta^r}_{ij})\boldsymbol{\jmath}$, the modeler will attribute a uniform distribution over 2π to the angles α_t, α_r,${\beta^t}_{ij}$ and ${\beta^r}_{ij}$.

With all these probability distributions derived and using the same methodology as in the narrowband (in terms of frequency selectivity) MIMO model proof, one can easily show that $\mathbf{\Theta}_{s_r \times s_t}$ i.i.d Gaussian is solution of the consistency argument and maximizes entropy.

Note that in the case $f = 0$, $\mathbf{v_t} = 0$ and $\mathbf{v_r} = 0$, the same model as the narrowband model is obtained. If more information is available on correlation or different variances of frequency paths, then this information can be incorporated in the matrix $\mathbf{D}_{s_r \times s_t}$, also known as the channel pattern mask [31]. Note that in the case of a ULA (Uniform Linear Array) geometry and in the Fourier directions, we have $\mathbf{u}^r_{ij} = \mathbf{u}^r_j$ (any column of

[19] Why does normality always appear in our models? Well, the answer is quite simple. In all this paper, we have always limited ourselves to the second moment of the channel. If more moments are available, then normal distributions would not appear in general.

matrix $\mathbf{\Phi}$ has a given direction) and $\mathbf{u}_{ij}^t = \mathbf{u}_i^t$ (any line of matrix $\mathbf{\Psi}$ has a given direction). Therefore, the channel model simplifies to:

$$\mathbf{H}(f,t) = \frac{1}{\sqrt{s_r s_t}} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ e^{j2\pi \frac{d(n_r-1)\sin(\phi_1)}{\lambda}} & \dots & e^{j2\pi \frac{d(n_r-1)\sin(\phi_{s_r})}{\lambda}} \end{pmatrix} \mathbf{\Theta}_{s_r \times s_t} \bigodot \mathbf{D}_{s_r \times s_t}(f,t)$$
$$\begin{pmatrix} 1 \dots & e^{j2\pi \frac{d(n_t-1)\sin(\psi_1)}{\lambda}} \\ \vdots \ddots & \vdots \\ 1 \dots & e^{j2\pi \frac{d(n_t-1)\sin(\psi_{s_t})}{\lambda}} \end{pmatrix}$$

In this case, the pattern mask $\mathbf{D}_{s_r \times s_t}$ has the following form:

$$\mathbf{D}_{s_r \times s_t}(f,t) = \begin{pmatrix} \sqrt{P_1^r}\sqrt{P_1^t} e^{-j2\pi f \tau_{1,1}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_1^r \mathbf{v}_r + \mathbf{u}_1^t \mathbf{v}_t)} & \dots & \sqrt{P_1^r}\sqrt{P_{s_t}^t} e^{-j2\pi f \tau_{1,s_t}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_1^r \mathbf{v}_r + \mathbf{u}_{s_t}^t \mathbf{v}_t)} \\ \vdots & \ddots & \vdots \\ \sqrt{P_{s_r}^r}\sqrt{P_1^t} e^{-j2\pi f \tau_{s_r,1}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_{s_r}^r \mathbf{v}_r + \mathbf{u}_1^t \mathbf{v}_t)} & \dots & \sqrt{P_{s_r}^r}\sqrt{P_{s_t}^t} e^{-j2\pi f \tau_{s_r,s_t}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_s^r \mathbf{v}_r + \mathbf{u}_{s_t}^t \mathbf{v}_t)} \end{pmatrix}$$

Although we take into account many parameters, the final model is quite simple. It is the product of three matrices: Matrices $\mathbf{\Phi}$ and $\mathbf{\Psi}$ taking into account the directions of arrival and departure; matrix $\mathbf{\Theta}_{s_r \times s_t} \bigodot \mathbf{D}_{s_r \times s_t}$ which is an independent Gaussian matrix with different variances. The frequency selectivity of the channel is therefore taken into account in the phase of each entry of the matrix $\mathbf{\Theta}_{s_r \times s_t} \bigodot \mathbf{D}_{s_r \times s_t}(f,t)$.

Remark: In the case of a one antenna system link ($n_r = 1$ and $n_t = 1$), we obtain:

$$\mathbf{H}(f,t) = \frac{1}{\sqrt{s_r s_t}} \left[e^{j(\phi_1 + 2\pi \frac{f \mathbf{u}_1^r \mathbf{v_r}}{c} t)} \dots e^{j(\phi_{s_r} + 2\pi \frac{f \mathbf{u}_{s_r}^r \mathbf{v_r}}{c} t)} \right] \begin{pmatrix} \sqrt{P_1^r} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sqrt{P_{s_r}^r} \end{pmatrix}$$
$$\mathbf{\Theta}_{s_r \times s_t} \bigodot \mathbf{D}_{s_r \times s_t}(f) \begin{pmatrix} \sqrt{P_1^t} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sqrt{P_{s_t}^t} \end{pmatrix} \begin{bmatrix} e^{j(\psi_1 + 2\pi \frac{f \mathbf{u}_1^t \mathbf{v_t}}{c} t)} \\ \vdots \\ e^{j(\psi_{s_t} + 2\pi \frac{f \mathbf{u}_{s_t}^t \mathbf{v_t}}{c} t)} \end{bmatrix}$$
$$= \frac{1}{\sqrt{s_r s_t}} \left[\sum_{k=1}^{s_r} \theta_{k,1} \sqrt{P_k^r} e^{j(\phi_k + 2\pi \frac{f \mathbf{u}_k^r \mathbf{v_r}}{c} t)} e^{-j2\pi f \tau_{k,1}} \dots \sum_{k=1}^{s_r} \theta_{k,s_R} \sqrt{P_k^r} e^{j(\phi_k + 2\pi \frac{f \mathbf{u}_k^r \mathbf{v_r}}{c} t)} e^{-j2\pi f \tau_{k,s_r}} \right]$$
$$\begin{pmatrix} \sqrt{P_1^t} & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \sqrt{P_{s_t}^t} \end{pmatrix} \begin{pmatrix} e^{j(\psi_1 + 2\pi \frac{f \mathbf{u}_1^t \mathbf{v_t}}{c} t)} \\ \vdots \\ e^{j(\psi_{s_t} + 2\pi \frac{f \mathbf{u}_{s_t}^t \mathbf{v_t}}{c} t)} \end{pmatrix}$$
$$= \sum_{l=1}^{s_t} \sum_{k=1}^{s_r} \rho_{k,l} e^{j2\pi \xi_{k,l} t} e^{-j2\pi f \tau_{k,l}}$$

where $\rho_{k,l}$ ($\rho_{k,l} = \frac{1}{\sqrt{s_r s_t}} \theta_{k,l} \sqrt{P_k^r} \sqrt{P_l^t} e^{j(\phi_k + \psi_l)}$) are independent Gaussian variable with zero mean and variance $\mathbb{E}(\mid \rho_{k,l} \mid^2) = \frac{1}{s_r s_t} P_k^r P_l^t$, $\xi_{k,l} = \frac{f}{c}(\mathbf{u}_k^r \mathbf{v_r} - \mathbf{u}_l^t \mathbf{v_t})$ are the doppler effect and $\tau_{k,l}$ are the delays. This previous result is a generalization of the SISO (Single Input Single Output) wireless model in the case of multifold scattering with the power profile taken into account.

5 Discussion

5.1 Müller's Model

In a paper "A Random Matrix Model of Communication via Antenna Arrays" [32], Müller develops a channel model based on the product of two random matrices:

$$\mathbf{H} = \mathbf{\Phi A \Theta}$$

where $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are two random matrices with zero mean unit variance i.i.d entries and $\mathbf{A}$ is a diagonal matrix (representing the attenuations). This model is intended to represent the fact that each signal bounces off a scattering object exactly once. $\mathbf{\Phi}$ represents the steering directions from the scatterers to the receiving antennas while $\mathbf{\Theta}$ represents the steering directions from the transmitting antennas to the scatterers. Measurements in [32] confirmed the model quite accurately. Should we conclude that signals in day to day life bounce only once on the scattering objects?

With the maximum entropy approach developed in this contribution, new insights can be given on this model and explanations can be provided on why Müller's model works so well. In the maximum entropy framework, Müller's model can be seen as either:

- a DoA based model with random directions i.e matrix $\mathbf{\Phi}$ with different powers (represented by matrix $\mathbf{A}$) for each angle of arrival. In fact, the signal can bounce freely several times from the transmitting antennas to the final scatterers (matrix $\mathbf{\Theta}$). Contrary to past belief, this model takes into account multi-fold scattering and answers the following question from a maximum entropy standpoint: what is the consistent model when the state of knowledge is limited to:
  - Random directions scattering at the receiving side.
  - Each steering vector at the receiving side has a certain power.
  - Each frequency path has a given variance.
- a corresponding DoD based model with random directions i.e matrix $\mathbf{\Theta}$ with different powers (represented by matrix $\mathbf{A}$) for each angle of departure. The model permits also in this case the signal to bounce several times from the scatterers to the receiving antennas. From a maximum entropy standpoint, the model answers the following question: what is the consistent model when the state of knowledge is limited to:
  - Random directions scattering at the transmitting side.
  - Each steering vector at the transmitting side has a certain power.
  - Each frequency has zero mean and a certain variance.
- DoA-DoD based model with random directions where the following question is answered: What is the consistent model when the state of knowledge is limited to:
  - Random directions scattering at the receiving side.
  - Random directions scattering at the transmitting side.
  - Each angle of arrival is linked to one angle of departure.

As one can see, Müller's model is broad enough to include several maximum entropy directional models and this fact explains why the model complies so accurately with the measurements performed in [33]

5.2 Sayeed's Model

In a paper "Deconstructing Multi-antenna Fading Channels" [34], Sayeed proposes a virtual representation of the channel. The model is the following:

$$\mathbf{H} = \mathbf{A}_{n_r}\mathbf{S}{\mathbf{A}_{n_t}}^H$$

Matrices $\mathbf{A}_{n_r}$ and $\mathbf{A}_{n_t}$ are discrete Fourier matrices and $\mathbf{S}$ is a $n_r \times n_t$ matrix which represents the contribution of each of the fixed DoA's and DoD's. The representation is virtual in the sense that it does not represent the real directions but only the contribution of the channel to those fixed directions. The model is somewhat a projection of the real steering directions onto a Fourier basis. Sayeed's model is quite appealing in terms of simplicity and analysis (it corresponds to the Maxent model on Fourier directions). In this case, also, we can revisit Sayeed's model in light of our framework. We can show that every time, Sayeed's model answers a specific question based on a given assumption.

- Suppose matrix $\mathbf{S}$ has i.i.d zero mean Gaussian entries then Sayeed's model answers the following question: what is the consistent model for a ULA when the modeler knows that the channel carries some energy, the DoA and DoD are on Fourier directions but one does not know what happens in between.
- Suppose now that matrix $\mathbf{S}$ has a certain correlation structure then Sayeed's model answers the following question: what is the consistent model for a ULA when the modeler knows that the channel carries some energy, the DoA and DoD are on Fourier directions but assumes that the paths in between have a certain correlation.

As one can see, Sayeed's model has a simple interpretation in the maximum entropy framework: it considers a ULA geometry with Fourier directions each time. Although it may seem strange that Sayeed

limits himself to Fourier directions, we do have an explanation for this fact. In his paper [31], Sayeed was mostly interested in the capacity scaling of MIMO channels and not the joint distribution of the elements. From that perspective, only the statistics of the uncorrelated scatterers is of interest since they are the ones which scale the mutual information. The correlated scatterers have very small effect on the information. In this respect, we must admit that Sayeed's intuition is quite impressive. However, the entropy framework is not limited to the ULA case (for which the Fourier vector approach is valid) and can be used for any kind of antenna and field approximation. One of the great features of the maximum entropy (which is not immediate in Sayeed's representation) approach is the quite simplicity for translating any additional physical information into probability assignment in the model. A one to one mapping between information and model representation is possible. With the maximum entropy approach, every new information on the environment can be straightforwardly incorporated and the models are consistent: adding or retrieving information takes us one step forward or back but always in a consistent way. The models are somewhat like Russian dolls, imbricated one into the other.

5.3 The "Kronecker" model

In a paper "Capacity Scaling in MIMO Wireless Systems Under Correlated fading", Chuah et al. study the following Kronecker [20] model:

$$\mathbf{H} = \mathbf{R}_{n_r}{}^{\frac{1}{2}} \mathbf{\Theta} \mathbf{R}_{n_t}{}^{\frac{1}{2}}$$

Here, $\mathbf{\Theta}$ is an $n_r \times n_t$ i.i.d zero mean Gaussian matrix, $\mathbf{R}_{n_r}{}^{\frac{1}{2}}$ is an $n_r \times n_r$ receiving correlation matrix while $\mathbf{R}_{n_t}{}^{\frac{1}{2}}$ is a $n_t \times n_t$ transmitting correlation matrix. The correlation is supposed to decrease sufficiently fast so that $\mathbf{R}_{n_r}$ and $\mathbf{R}_{n_t}$ have a Toeplitz band structure. Using a software tool (Wireless System Engineering [37]), they demonstrate the validity of the model. Quite remarkably, although designed to take into account receiving and transmitting correlation, the model developed in the paper falls within the double directional framework. Indeed, since $\mathbf{R}_{n_r}$ and $\mathbf{R}_{n_t}$ are band Toeplitz then these matrices are asymptotically diagonalized in a Fourier basis

$$\mathbf{R}_{n_r} \sim F_{n_r} \Lambda_{n_r} F_{n_r}^H$$

and

$$\mathbf{R}_{n_t} \sim F_{n_t} \Lambda_{n_t} F_{n_t}^H.$$

F_{n_r} and F_{n_t} are Fourier matrices while Λ_{n_r} and Λ_{n_t} represent the eigenvalue matrices of $\mathbf{R}_{n_r}$ and $\mathbf{R}_{n_t}$.

Therefore, matrix $\mathbf{H}$ can be rewritten as:

$$\begin{aligned}
\mathbf{H} &= \mathbf{R}_{n_r}{}^{\frac{1}{2}} \mathbf{\Theta} \mathbf{R}_{n_t}{}^{\frac{1}{2}} \\
&= F_{n_r} \left(\Lambda_{n_r}{}^{\frac{1}{2}} F_{n_r}{}^H \mathbf{\Theta} F_{n_t} \Lambda_{n_t}{}^{\frac{1}{2}} \right) F_{n_t}{}^H \\
&= F_{n_r} \left(\mathbf{\Theta_1} \bigodot \mathbf{D}_{n_r \times n_t} \right) F_{n_t}{}^H
\end{aligned}$$

$\mathbf{\Theta_1} = \mathbf{F}_{n_r}{}^H \mathbf{\Theta} \mathbf{F}_{n_t}$ is a $n_r \times n_t$ zero mean i.i.d Gaussian matrix and $\mathbf{D}_{n_r \times n_t}$ is a pattern mask matrix defined by:

$$\mathbf{D}_{s \times s_1} = \begin{pmatrix} \lambda_{1,n_t}^{\frac{1}{2}} \lambda_{1,n_r}^{\frac{1}{2}} & \dots & \lambda_{n_t,n_t}^{\frac{1}{2}} \lambda_{1,n_r}^{\frac{1}{2}} \\ \vdots & \ddots & \vdots \\ \lambda_{1,n_t}^{\frac{1}{2}} \lambda_{n_r,n_r}^{\frac{1}{2}} & \dots & \lambda_{n_t,n_t}^{\frac{1}{2}} \lambda_{n_r,n_r}^{\frac{1}{2}} \end{pmatrix}$$

Note that this connection with the double directional model has already been reported in [31]. Here again, the previous model can be reinterpreted in light of the maximum entropy approach. The model answers the following question: what is the consistent model one can make when the DoA are uncorrelated and have respective power λ_{i,n_r}, the DoD are uncorrelated and have respective power λ_{i,n_t}, each path has zero mean and a certain variance. The model therefore confirms the double directional assumption as well as Sayeed's approach and is a particular case of the maximum entropy approach. The comments and limitations made on Sayeed's model are also valid here. **reference also [38, 39]**

[20] The model is called a Kronecker model because $\mathbb{E}(\mathrm{vec}(\mathbf{H})^H \mathrm{vec}(\mathbf{H})) = \mathbf{R}_{n_r} \bigotimes \mathbf{R}_{n_t}$ is a Kronecker product. The justification of this approach relies on the fact that only immediate surroundings of the antenna array impose the correlation between array elements and have no impact on correlations observed between the elements of the array at the other end of the link. Some discussions can be found in [35, 36].

5.4 The "Keyhole" Model

In [40], Gesbert et al. show that low correlation[21] is not a guarantee of high capacity: cases where the channel is rank deficient can appear while having uncorrelated entries (for example when a screen with a small keyhole is placed in between the transmitting and receiving antennas). In [42], they propose the following model for a rank one channel:

$$\mathbf{H} = \mathbf{R}_{n_r}{}^{\frac{1}{2}} \mathbf{g}_r \mathbf{g}_t{}^H \mathbf{R}_{n_t}{}^{\frac{1}{2}} \tag{8}$$

Here, $\mathbf{R}_{n_r}{}^{\frac{1}{2}}$ is an $n_r \times n_r$ receiving correlation matrix while $\mathbf{R}_{n_t}{}^{\frac{1}{2}}$ is a $n_t \times n_t$ transmitting correlation matrix. $\mathbf{g}_r$ and $\mathbf{g}_t$ are two independent transmit and receiving Rayleigh fading vectors. Here again, this model has connections with the previous maximum entropy model:

$$\mathbf{H} = \frac{1}{\sqrt{s_r s_t}} \mathbf{\Phi}_{n_r \times s_r} \mathbf{\Theta}_{s_r \times s_t} \mathbf{\Psi}_{s_t \times n_t} \tag{9}$$

The Keyhole model can be either:

- A double direction model with $s_r = 1$ and $\mathbf{\Phi}_{n_r \times 1} = \mathbf{R}_{n_r}{}^{\frac{1}{2}} \mathbf{g}_r$. In this case, $\mathbf{g}_t{}^H \mathbf{R}_{n_t}{}^{\frac{1}{2}} = \mathbf{\Theta}_{1 \times s_t} \mathbf{\Psi}_{s_t \times n_t}$ where $\mathbf{\Theta}_{1 \times s_t}$ is zero mean i.i.d Gaussian.
- A double direction model with $s_t = 1$ and $\mathbf{\Psi}_{1 \times n_t} = \mathbf{g}_t{}^H \mathbf{R}_{n_t}{}^{\frac{1}{2}}$. In this case, $\mathbf{R}_{n_r}{}^{\frac{1}{2}} \mathbf{g}_r = \Phi_{n_r \times s_r} \mathbf{\Theta}_{s_r \times 1}$ where $\mathbf{\Theta}_{s_r \times 1}$ is zero mean i.i.d Gaussian.

As one can observe, the maximum entropy model can take into account rank deficient channels.

5.5 Conclusion

After analyzing each of these models, we find that they all answer a specific question based on a given state of knowledge. All these models can be derived within the maximum entropy framework and have a simple interpretation. Moreover, each time the directional assumption appears which conjectures the correctness of the directional approach.

6 Testing the Models

In all the previous sections, we have developed several models based on different questions. But what is the right model, in other words how to choose between the set $\{M_0, M_1, ..., M_K\}$ of K models (note that M specifies only the type of model and not the parameters of the model)?

6.1 Bayesian Viewpoint

When judging the appropriateness of a model, Bayes[22] rules derives the posterior probability of the model. Bayes rule gives the posterior probability for the i^{th} model according to: [23]

$$P(M_i \mid Y, I) = P(M_i \mid I) \frac{P(Y \mid M_i, I)}{P(Y \mid I)}$$

Y is the data (given by measurements), I is the prior information (ULA, far field scattering...). For comparing two models M and M_1, one has to compute the ratio:

[21] "keyhole" channels are MIMO channels with uncorrelated spatial fading at the transmitter and the receiver but have a reduced channel rank (also known as uncorrelated low rank models). They were shown to arise in roof-edge diffraction scenarios [41].

[22] This chapter is greatly inspired by the work of Jaynes and Bretthorst who have made the following ideas clear.

[23] We use here the notations and meanings of Jaynes [20] and Jeffrey [18]: $P(M_i \mid Y, I)$ is the "probability that the model M_i is true given that the data Y is equal to the true data y and that the information I on which is based the model is true". Every time, " (|" means conditional on the truth of the hypothesis I. In probability theory, all probabilities are conditional on some hypothesis space.

$$\frac{P(M_1 \mid Y,I)}{P(M \mid Y,I)} = \frac{P(M_1 \mid I)}{P(M \mid I)} \frac{P(Y \mid M_1,I)}{P(Y \mid M,I)}$$

If $P(M_1 \mid Y,I) > P(M \mid Y,I)$, then one will conclude that model M_1 is better than model M. Let us now try to understand each term.
The first term, crucially important, is usually forgotten by the channel modelling community: $\frac{P(M_1|I)}{P(M|I)}$. It favors one model or the other before the observation. As an example, suppose that the information $\{I =$ The scatterers are near the antennas $\}$ is given. Then if one has to compare the model M (which considers ULA with far field scattering) and the model M_1 (assuming near field scattering) then one should consider $\frac{P(M_1|I)}{P(M|I)} > 1$. [24]

For understanding the second term, let us analyze and compare the following two specific models: the DoA based model M_{doa} and the double directional model M_{double}.

Model M_{doa}:

$$\mathbf{H}(f,t) = \frac{1}{\sqrt{s_r}} \Phi \left(\mathbf{\Theta} \bigodot D(t,f) \right)$$

with

$$D(t,f) = \begin{pmatrix} e^{-j2\pi f \tau_{1,1}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_1^r \mathbf{v}_r)} & \dots & e^{-j2\pi f \tau_{1,n_t}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_1^r \mathbf{v}_r)} \\ \vdots & \ddots & \vdots \\ e^{-j2\pi f \tau_{s_r,1}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_s^r \mathbf{v}_r)} & \dots & e^{-j2\pi f \tau_{s_r,n_t}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_s^r \mathbf{v}_r)} \end{pmatrix}$$

deals with the DoA model taking into account the delays, Doppler effect (we suppose that the transmitting antenna does not move but only the receiving one) for a ULA (s is the number of scatterers). Let the information I on which is based the model be such that the powers of the steering directions are identical and that the transmitting antennas do not move. We recall that $\mathbf{u}_i^r \mathbf{v}_r = (\cos({\beta^r}_i)\mathbf{i} + \sin({\beta^r}_i)\mathbf{j})(v_r \cos(\alpha_r)\mathbf{i} + v_r \sin(\alpha_r)\mathbf{j}) = v_r \cos({\beta^r}_i - \alpha_r)$

The set of parameters on which the model is defined is

$$p_{\text{doa}} = \{\mathbf{\Phi}, s_r, \tau, v_r, \mathbf{\Theta}, \alpha_r, \beta^r\}$$

and the parameters lie in a subspace $S_{p_{\text{doa}}}$. We recall here the DoA based model for a given frequency:

$$\mathbf{y}(t,f) = \frac{1}{\sqrt{s_r}} \mathbf{\Phi} \left(\mathbf{\Theta} \bigodot D(t,f) \right) \mathbf{x}(f) + \mathbf{n}(f)$$

The term of interest $P(\mathbf{y} \mid M_{\text{doa}}, I)$ can be derived the following way:

$$P(\mathbf{y} \mid M_{\text{doa}}, I) = \int P(\mathbf{y}, p_{\text{doa}} \mid M_{\text{doa}}, I) dp_{\text{doa}} = \int P(\mathbf{y} \mid p_{\text{doa}}, M_{\text{doa}}, I) P(p_{\text{doa}} \mid M_{\text{doa}}, I) dp_{\text{doa}}$$

Let us derive each probability distribution separately: $P(\mathbf{y} \mid p_{\text{doa}}, M_{\text{doa}}, I) =$

$$\frac{1}{(2\pi\sigma^2)^{\frac{N_1 N_r}{2}}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N_1} \left(y(t_j,f_i) - \frac{1}{\sqrt{s_r}} \mathbf{\Phi}(\mathbf{\Theta} \odot D(t_j,f_i)) x(f_i) \right)^H \left(y(t_j,f_i) - \frac{1}{\sqrt{s_r}} \mathbf{\Phi}(\mathbf{\Theta} \odot D(t_j,f_i)) x(f_i) \right)}$$

and

$$\begin{aligned} P(p_{\text{doa}} \mid M_{\text{doa}}, I) &= P(\mathbf{\Phi}, s_r, \tau, \beta^{\mathbf{r}}, v_r, \alpha_r, \mathbf{\Theta} \mid M_{\text{doa}}, I) \\ &= P(\mathbf{\Phi} \mid s_r, M_{\text{doa}}, I) P(s_r \mid M_{\text{doa}}, I) P(v_r \mid M_{\text{doa}}, I) P(\tau \mid M_{\text{doa}}, I) \\ &\quad P(\mathbf{\Theta} \mid M_{\text{doa}}, s_r, I) P(\alpha_r \mid M_{\text{doa}}, I) P(\beta^r \mid I, M_{\text{doa}}) \end{aligned}$$

since all the priors are taken independent in the case of uninformative priors. The values of these priors have already been provided (the proof is given in chapter 4.3) and only the prior on $\mathbf{\Theta}$ and s_r remain to be given. We give these two priors now (and also the prior on the power although in the two models introduced for comparison, the power distribution is not needed):

[24] The term $\frac{P(M_1|I)}{P(M|I)}$ can be seen as the revenge of the measurement field scientist over the mathematician. It shows that modelling is both an experimental and theoretical science and that the experience of the field scientist (which attributes the values of the prior probabilities) does matter.

- If only the mean and variance of each path is available then using maximum entropy arguments, one can show that:

$$P(\boldsymbol{\Theta} \mid s_r, M_{\text{doa}}, I) = \frac{1}{(\sqrt{2\pi})^{n_t \times s_r}} e^{-\sum_{i=1}^{s_r}\sum_{j=1}^{n_t}|\theta_{i,j}|^2} = \frac{1}{(\sqrt{2\pi})^{n_t \times s_r}} e^{-\text{trace}(\boldsymbol{\Theta}\boldsymbol{\Theta}^H)}$$

- How can we assign a prior probability $P(s_r \mid M_{\text{doa}}, I)$ for the unknown number of scatterers? The modeler has no knowledge if the measurements were taken in a dense area or not. The unknown number of scatterers could range from one (this prior only occurs in model that have a single bounce) up to a maximum. But what is the maximum value? There are $N \times N_1$ data values and if there were $N \times N_1$ scatterers, the data could be at most fit by placing a scatterer at each data value and adjusting the direction of arrivals. Because no additional information is available about the number of scatterers, $N \times N_1$ may be taken as an upper bound. Using the principle of maximum entropy, one obtains a uniform distribution for the number of scatterers $P(s_r \mid M_{\text{doa}}, I) = \frac{1}{N \times N_1}$.
  Note that in the general case, if one has precise available information then one has to take it into account. But how can the modeler translate the prior on the scatterers due to the fact that the room has three chairs and a lamp in the corner? This is undoubtedly a difficult task and representing that information in terms of probabilities is not straightforward. But difficult is not impossible. The fact that there are several chairs (with respect to the case where there is no chairs) is a source of information and will lead to attributing in the latter case a peaky prior shifted around a higher number of scatterers.
- **Power**: The transmitter is limited in terms of transmit power to an upper bound value $P^t_{\max}$. Therefore, the principle of maximum entropy attributes a uniform distribution to the different amplitudes $P({P_i}^t) = \frac{1}{P^t_{\max}}$, $P_i \in [0, P^t_{\max}]$. In the same vein, the receiver cannot, due to the amplifiers, process a receiving amplitude greater then $P^r_{\max}$. In this case, the principle of maximum entropy attributes a uniform distribution such as $P(P^r_i) = \frac{1}{P^r_{\max}}$, $P_i \in [0, P^r_{\max}]$

  With all the previous priors given, one can therefore compute:

$$\begin{aligned} P(\mathbf{y} \mid M_{\text{doa}}, I) = \int & \frac{1}{(2\pi\sigma^2)^{\frac{N_1 N_r}{2}}} \\ & e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{N_1}\left(y(t_j,f_i)-\frac{1}{\sqrt{s_r}}\boldsymbol{\Phi}\left(\boldsymbol{\Theta}\odot D(t_j,f_i)\right)x(f_i)\right)^H\left(y(t_j,f_i)-\frac{1}{\sqrt{s_r}}\boldsymbol{\Phi}\left(\boldsymbol{\Theta}\odot D(t_j,f_i)\right)x(f_i)\right)} \\ & P(\boldsymbol{\Phi} \mid s_r, M_{\text{doa}}, I) P(s_r \mid M_{\text{doa}}, I) P(v_r \mid M_{\text{doa}}, I) P(\alpha_r \mid M_{\text{doa}}, I) P(\beta^r \mid M_{\text{doa}}, I) \\ & P(\tau \mid M_{\text{doa}}, I) P(\boldsymbol{\Theta} \mid M_{\text{doa}}, I) d\boldsymbol{\Phi} d\boldsymbol{\Theta} ds_r d\tau dv_r d\alpha_r d\beta^{\mathbf{r}} \end{aligned}$$

which gives:

$$\begin{aligned} P(\mathbf{y} \mid M_{\text{doa}}, I) = & \frac{1}{N \times N_1} \sum_{s_r=1}^{N \times N_1} \int_0^{2\pi}\int_0^{\infty}\int_0^{v_{\lim}}\int_0^{\tau_{\max}} \frac{1}{(2\pi\sigma^2)^{\frac{N_1 N_r}{2}}} \prod_{i=1}^{N}\prod_{j=1}^{N_1} \\ & e^{-\frac{1}{2\sigma^2}\left(y(t_j,f_i)-\frac{1}{\sqrt{s_r}}\boldsymbol{\Phi}(\boldsymbol{\Theta}\odot D(t_j,f_i))x(f_i)\right)^H\left(y(t_j,f_i)-\frac{1}{\sqrt{s_r}}\boldsymbol{\Phi}(\boldsymbol{\Theta}\odot D(t_j,f_i))x(f_i)\right)} \\ & \left(\frac{1}{\tau_{\max}}\right)^{s_r \times n_t} \frac{1}{v_{\lim}} \left(\frac{1}{2\pi}\right)^{n_r \times s_r} \frac{1}{2\pi}\left(\frac{1}{2\pi}\right)^{s_r} \\ & d\phi_{11}...d\phi_{n_r s_r} d\theta_{11}...d\theta_{s_r n_t} d\tau_{11}...d\tau_{s_r n_t} dv_r d\alpha_r d{\beta^r}_1 ...d{\beta^r}_{s_r} \end{aligned} \tag{10}$$

As one can see, the numerical integration is tedious but it is the only way to rank the models in an appropriate manner.

Model M_{double}:

Let us now derive model M_{double}:

$$\mathbf{H}(f,t) = \frac{1}{\sqrt{s_r s_t}} \boldsymbol{\Phi}\left(\boldsymbol{\Theta} \bigodot D(t,f)\right)\boldsymbol{\Psi}$$

with

$$D(t,f) = \begin{pmatrix} e^{-j2\pi f\tau_{1,1}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_1^r \mathbf{v}_r)} & \dots & e^{-j2\pi f\tau_{1,s_t}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_1^r \mathbf{v}_r)} \\ \vdots & \ddots & \vdots \\ e^{-j2\pi f\tau_{s_r,1}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_{s_r}^r \mathbf{v}_r)} & \dots & e^{-j2\pi f\tau_{s_r,s_t}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_{s_r}^r \mathbf{v}_r)} \end{pmatrix}$$

deals with the double directional model for which the set of parameters is

$$p_{\text{double}} = \{\mathbf{\Phi}, s_r, \mathbf{\Psi}, s_t, \tau, v_r, , \alpha_r, \beta_{\mathbf{r}} \mathbf{\Theta}\} = \{p_{\text{doa}}, \mathbf{\Phi}, s_t\}$$

by adding two new parameters $\mathbf{\Psi}$ and s_t and going to the new subspace $S_{p_{\text{double}}}$ in such a way that $\mathbf{\Psi} = \mathbf{F}_{n_t}$ ($n_t = s_t$) represents model M_{doa}. Indeed, in this case, we have:

$$\begin{aligned} \left(\mathbf{\Theta} \bigodot D(t,f)\right) \mathbf{F}_{n_t} &= \begin{pmatrix} \sum_{i=1}^{n_t} \theta_{1i} e^{-j2\pi f\tau_{1,i}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_1^r \mathbf{v}_r)} & \dots & \sum_{i=1}^{n_t} \theta_{1i} e^{-j2\pi f\tau_{1,i}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_1^r \mathbf{v}_r)} e^{j2\pi \frac{(n_t-1)i}{n_t}} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n_t} \theta_{s_r i} e^{-j2\pi f\tau_{s_r,i}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_s^r \mathbf{v}_r)} & \dots & \sum_{i=1}^{n_t} \theta_{s_r i} e^{-j2\pi f\tau_{s_r,i}} e^{j2\pi \frac{ft}{c}(\mathbf{u}_{s_r}^r \mathbf{v}_r)} e^{j2\pi \frac{(n_t-1)i}{n_t}} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n_t} \theta_{1i} e^{-j2\pi f(\tau_{1,i}-\tau_{1,1})} & \dots & \sum_{i=1}^{n_t} \theta_{1i} e^{-j2\pi f(\tau_{1,i}-\tau_{1,n_t})} e^{j2\pi \frac{(n_t-1)i}{n_t}} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n_t} \theta_{s_r i} e^{-j2\pi f(\tau_{s_r,i}-\tau_{s_r,1})} & \dots & \sum_{i=1}^{n_t} \theta_{s_r i} e^{-j2\pi f(\tau_{s_r,i}-\tau_{s_r,n_t})} e^{j2\pi \frac{(n_t-1)i}{n_t}} \end{pmatrix} \bigodot D(t,f) \\ &= \Theta_1 \bigodot D(t,f) \end{aligned}$$

Where Θ_1 is a matrix with i.i.d Gaussian entries.
We recall here the model for a given frequency:

$$\mathbf{y}(f,t) = \frac{1}{\sqrt{s_r s_t}} \mathbf{\Phi} \left(\mathbf{\Theta} \bigodot D(t,f)\right) \mathbf{\Psi} \mathbf{x}(f) + \mathbf{n}(f)$$

The same methodology applies and we have:

$$\begin{aligned} P(\mathbf{y} \mid M_{\text{double}}, I) = \int & \frac{1}{(2\pi\sigma^2)^{\frac{N_1 N_r}{2}}} \\ & e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{j=1}^{N_1} \left(y(t_j,f_i) - \frac{1}{\sqrt{s_r s_t}} \mathbf{\Phi}\left(\mathbf{\Theta} \odot D(t_j,f_i)\right) \mathbf{\Psi} x(f_i)\right)^H \left(y(t_j,f_i) - \frac{1}{\sqrt{s_r s_t}} \mathbf{\Phi}\left(\mathbf{\Theta} \odot D(t_j,f_i)\right) \mathbf{\Psi} x(f_i)\right)} \\ & P(\mathbf{\Phi} \mid s_r, M_{\text{double}}, I) P(s_r \mid M_{\text{double}}, I) P(\mathbf{\Psi} \mid s_t, M_{\text{double}}, I) P(s_t \mid M_{\text{double}}, I) \\ & P(v_r \mid M_{\text{double}}, I) P(\alpha_r \mid M_{\text{double}}, I) P(\beta_{\mathbf{r}} \mid M_{\text{double}}, I) P(\tau \mid M_{\text{double}}, I) \\ & P(\mathbf{\Theta} \mid M_{\text{double}}, I) d\mathbf{\Phi} d\mathbf{\Psi} d\mathbf{\Theta} ds_r ds_t d\tau dv_r d\alpha_r d\beta^{\mathbf{r}} \end{aligned}$$

and

$$\begin{aligned} P(\mathbf{y} \mid M_{\text{double}}, I) = & \left(\frac{2}{N \times N_1}\right)^2 \sum_{s_r=1}^{\frac{N \times N_1}{2}} \sum_{s_t=1}^{\frac{N \times N_1}{2}} \int_0^{2\pi} \int_0^{\infty} \int_0^{v_{\lim}} \int_0^{\tau_{\max}} \frac{1}{(2\pi\sigma^2)^{\frac{N_1 N_r}{2}}} \prod_{i=1}^{N} \prod_{j=1}^{N_1} \\ & e^{-\frac{1}{2\sigma^2} \left(y(t_j,f_i) - \frac{1}{\sqrt{s_r s_t}} \mathbf{\Phi}(\mathbf{\Theta} \odot D(t_j,f_i)) \mathbf{\Psi} x(f_i)\right)^H \left(y(t_j,f_i) - \frac{1}{\sqrt{s_r s_t}} \mathbf{\Phi}(\mathbf{\Theta} \odot D(t_j,f_i)) \mathbf{\Psi} x(f_i)\right)} \\ & \left(\frac{1}{\tau_{\max}}\right)^{s_r \times s_t} \frac{1}{v_{\lim}} \left(\frac{1}{2\pi}\right)^{n_r \times s_r} \left(\frac{1}{2\pi}\right)^{s_t \times n_t} \frac{1}{2\pi} \left(\frac{1}{2\pi}\right)^{s_r} \\ & d\phi_{11} ... d\phi_{n_r s_r} d\psi_{11} ... d\psi_{1 n_t} d\theta_{11} ... d\theta_{s_r n_t} d\tau_{11} ... d\tau_{s_r n_t} dv_r d\alpha_r d\beta^r{}_1 ... d\beta^r{}_{s_r} \end{aligned} \quad (11)$$

A common problem in the modelling process is the following: suppose, when testing the models with the data, that both models M and M_1 have the same maximum likelihood, in other words:

$$P(\mathbf{y} \mid p_{\text{doa}}{}^{\max}, M_{\text{doa}}, I) = P(\mathbf{y} \mid p_{\text{double}}{}^{\max}, M_{\text{double}}, I)$$

Which model should we choose? Hereafter, we give an example to show that Bayesian probability will choose the model with the smallest number of parameters.

First of all, we will suppose that the information I available does not give a preference to model before seeing the data: $P(M_{\text{double}} \mid I) = P(M_{\text{doa}} \mid I)$.
As previously shown,

$$P(\mathbf{y} \mid M_{\text{doa}}, I) = \int P(\mathbf{y}, p_{\text{doa}} \mid M_{\text{doa}}, I) dp_{\text{doa}}$$
$$= \int P(\mathbf{y} \mid p_{\text{doa}}, M_{\text{doa}}, I) P(p_{\text{doa}} \mid M_{\text{doa}}, I) dp_{\text{doa}}$$

and

$$P(\mathbf{y} \mid M_{\text{double}}, I) = \int P(\mathbf{y}, p_{\text{double}} \mid M_{\text{double}}, I) dp_{\text{double}} \tag{12}$$
$$= \int P(\mathbf{y} \mid p_{\text{double}}, M_{\text{double}}, I) P(p_{\text{double}} \mid M_{\text{double}}, I) dp_{\text{double}} \tag{13}$$

Since

$$P(p_{\text{double}} \mid M_{\text{double}}, I) = P([p_{\text{doa}}, \mathbf{\Psi}, s_t] \mid M_{\text{double}}, I)$$
$$= P(p_{\text{doa}} \mid \mathbf{\Psi}, s_t, M_{\text{double}}, I) P(\mathbf{\Psi}, s_t \mid M_{\text{double}}, I)$$

From equation (12), we have:

$$P(\mathbf{y} \mid M_{\text{double}}, I) = \int\int P(\mathbf{y} \mid [p_{\text{doa}}, \mathbf{\Psi}, s_t], M_{\text{double}}, I) P(p_{\text{doa}} \mid \mathbf{\Psi}, s_t, M_{\text{double}}, I)$$
$$P(\mathbf{\Psi}, s_t \mid M_{\text{double}}, I) dp_{\text{doa}} d\mathbf{\Psi} ds_t$$

In the following, we will suppose that the likelihood function $P(\mathbf{y} \mid [p_{\text{doa}}, \mathbf{\Psi}, s_t], M_{\text{double}}, I)$ is peaky around the maximum likelihood region and has near zero values elsewhere. Otherwise, the measurement data Y would be useless in the sense that the data does not provide any information. Suppose now that with model M_{double}, the maximum likelihood $P(\mathbf{y} \mid [p_{\text{doa}}, \mathbf{\Psi}, s_t] M_{\text{double}}, I)$ occurs at a point near $\mathbf{\Psi} = \mathbf{F}_t$ and $s_t = n_t$ for the parameters $\mathbf{\Psi}$ and s_t in other words $P(\mathbf{y} \mid [p_{\text{doa}}, \mathbf{\Psi}, s_t], M_{\text{double}}, I)$ is always null except for the value of $\mathbf{\Psi} = \mathbf{F}_t$ and $s_t = n_t$ then:

$$P(\mathbf{y} \mid M_{\text{double}}, I) = \int\int\int P(\mathbf{y} \mid [p_{\text{doa}}, \mathbf{\Psi}, s_t], M_{\text{double}}, I) P(p_{\text{doa}} \mid \mathbf{\Psi}, s_t M_{\text{double}}, I)$$
$$P(\mathbf{\Psi}, s_t \mid M_{\text{double}}, I) dp_{\text{doa}} d\mathbf{\Psi} ds_t$$
$$\approx \int P(\mathbf{y} \mid [p_{\text{doa}}, \mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t], M_{\text{Double}}, I)$$
$$P(p_{\text{doa}} \mid [\mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t] M_{\text{double}}, I) \tag{14}$$
$$P([\mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t] \mid M_{\text{double}}, I) dp_{\text{doa}} \tag{15}$$

One has to notice that $P(p_{\text{doa}} \mid [\mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t], M_{\text{double}}, I) = P(p_{\text{doa}} \mid M_{\text{doa}}, I)$ and $P(\mathbf{y} \mid [p_{\text{doa}}, \mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t], M_{\text{double}}, I) = P(\mathbf{y} \mid p_{\text{doa}}, M_{\text{doa}}, I)$ since both models are the same when $\mathbf{\Psi} = \mathbf{F}_{n_t}$ and $s_t = n_t$. We also have $P(\mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t \mid M_{\text{double}}, I) \leq 1$ (In fact, we can derive the exact value. Indeed, since we have no knowledge of the directions of arrival, $P(\mathbf{\Phi} = \mathbf{F}_{n_t}, s_t = n_t \mid M_{\text{double}}, I) = \frac{1}{(2\pi)^{n_r \times s_t}}$). Using equation (14), Bayesian probability shows us that:

$$P(\mathbf{y} \mid M_{\text{double}}, I) \leq \int P(\mathbf{y} \mid [p_{\text{doa}}, \mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t], M_{\text{Double}}, I)$$
$$P(p_{\text{doa}} \mid [\mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t] M_{\text{double}}, I) P([\mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t] \mid M_{\text{double}}, I) dp_{\text{doa}}$$
$$= \int P(\mathbf{y} \mid p_{\text{doa}}, M_{\text{doa}}, I) P(p_{\text{doa}} \mid M_{\text{doa}}, I) P([\mathbf{\Psi} = \mathbf{F}_{n_t}, s_t = n_t] \mid M_{\text{double}}, I) dp_{\text{doa}}$$
$$\leq \int P(\mathbf{y} \mid p_{\text{doa}}, M_{\text{doa}}, I) P(p_{\text{doa}} \mid M_{\text{doa}}, I) dp_{\text{doa}}$$
$$= \int P(\mathbf{y}, p_{\text{doa}} \mid M_{\text{doa}}, I) dp_{\text{doa}}$$
$$P(\mathbf{y} \mid M_{\text{doa}}, I)$$

Since M_{doa} has less parameters then M_{double}, Bayesian probability will favor the model M_{doa} with less parameters and therefore shows that "the best explanation is always the simplest"[25]. It is therefore wrong

[25] In statistical inference, this is known as Occam's razor. William of Occam was a theologian of the 14th century who wrote against the papacy in a series of treatise in which he tried to avoid many established pseudo explanations. In

to think that by increasing the number of parameters one can always find a good model: one can indeed better fit the data to the model (expression $P(\mathbf{y} \mid p_{\text{doa}}, M_{\text{doa}}, I)$) but the prior probability $P(p_{\text{doa}} \mid M_{\text{doa}}, I)$ will spread over a larger space and assign as a consequence a lower value to $P(\mathbf{y} \mid M_{\text{doa}}, I)$.

But how does the a posteriori computation compare with the usual methodology of maximizing the likelihood $P(\mathbf{y} \mid p, M, I)$?

Following [20], let us expand $\log P(\mathbf{y} \mid p, M, I)$ around the maximum likelihood point $\hat{p} = \{{p^1}_{\max}, ..., {p^m}_{\max}\}$

$$\log P(\mathbf{y} \mid p, M, I) = \log P(\mathbf{y} \mid p_{\max}, M, I) + \frac{1}{2} \sum_{i,j=1}^{m} \frac{d^2 \log(P)}{dp^i dp^j} (p^i - {p^i}_{\max})(p^j - {p^j}_{\max}) + O()$$

then near the peak a good approximation is a multivariate Gaussian such as:

$$P(\mathbf{y} \mid p, M, I) = P(\mathbf{y} \mid p_{\max}, M, I) e^{-\frac{1}{2}(p-p_{\max})\Delta^{-1}(p-p_{\max})}$$

with the inverse covariance matrix defined as:

$${\Delta^{-1}}_{ij} = \left(\frac{d^2 \log(P)}{dp^i dp^j} \right)_{\pi=\pi_{\max}}$$

Therefore,

$$\begin{aligned} P(\mathbf{y} \mid M, I) &= P(\mathbf{y} \mid p_{\max}, M, I) \int e^{-\frac{1}{2}(p-p_{\max})\Delta^{-1}(p-p_{\max})} P(p \mid M, I) dp \\ &= P(\mathbf{y} \mid p_{\max}, M, I) G(M, I) \end{aligned}$$

All the tools are now provided to better understand what is happening. Suppose we want to compare two models M and M_1. The a posteriori probability ratio for model M over M_1 is:

$$\begin{aligned} \frac{P(M \mid \mathbf{y}, I)}{P(M_1 \mid \mathbf{y}, I)} &= \frac{P(M \mid I)}{P(M_1 \mid I)} \frac{P(\mathbf{y} \mid M, I)}{P(\mathbf{y} \mid M_1, I)} \\ &= \frac{P(M \mid I)}{P(M_1 \mid I)} \frac{P(\mathbf{y} \mid p^{\max}, M, I)}{P(\mathbf{y} \mid {p_1}^{\max}, M_1, I)} \frac{G(M, I)}{G(M_1, I)} \end{aligned}$$

In the conventional methods, M is better than M_1 if $\frac{P(\mathbf{y}|p^{\max} M, I)}{P(\mathbf{y}|{p_1}^{\max} M_1, I)} > 1$ which is only one part of the three terms to be computed. In fact, in order to compare two models, three terms have to be calculated and the mistake persists thinking that any model M_1 versus M is good as long as we increase the number of parameters: indeed, the fitting will get better and the ratio $\frac{P(\mathbf{y}|p^{\max}, M, I)}{P(\mathbf{y}|{p_1}_{\max}, M_1, I)}$ will decrease but this is only looking at one part of the problem. First of all, one has to consider $\frac{P(M|I)}{P(M_1|I)}$ and moreover $\frac{G(M,I)}{G(M_1,I)}$. This last term depends on the prior information about the internal parameters and as the number of parameters increases this term decreases due to the fact that we add more and more uninformative priors.

6.2 Conventional Methods

In the previous section, we have shown how probability theory can be used to rank the models. However, the integrals derived in equation (10) and equation (11) are not easy to compute, especially in the case of interest with a high number of antennas (8×8) since we have to marginalize our integrals across a great number of parameters. But however difficult the problem may be, it is not a reason to hide problems and the use of other methods should be clearly explained. The reader must now know that one can rank models and that there is an optimum number of parameters when representing information. The Bayesian framework gives us an answer by comparing the a posteriori probability ratios: $\frac{P(M|\mathbf{y},I)}{P(M_1|\mathbf{y},I)}$. If one is to use other testing methods, then one has to clearly understand the limitations of these methods and justify the use of the criteria. In the following, we explain two procedures used by the channel modelling community and explain their limitations.

---

his terms, the logic of simplicity was stated in the following form "Causes shall not be multiplied beyond necessity" [28]. Note that Occam's razor has been extended to other fields such as metaphysics where it is interpreted as "nature prefers simplicity".

1- Parameter estimation methods

In this procedure, the data is cut into two parts, one for estimating the parameters, the other to validate the model incorporating the parameters.

- For estimating the parameters such as the angles of arrival, non-parametric methods such as the beamforming or the Capon method [43] can be used. In the case of parametric methods such as Music [44], Min-Norm [45] or Esprit method [46], they rely on properties of the structure of the covariance $\mathbf{R} = \mathbb{E}(\mathbf{y}\mathbf{y}^{\mathbf{H}}) = \mathbf{\Phi}\mathbf{K}\mathbf{\Phi}^{\mathbf{H}} + \sigma^{\mathbf{2}}\mathbf{I}$ of the output signal. In this case, one has to assume that matrix $\mathbf{K}$ ($\mathbf{K} = \mathbb{E}(\mathbf{\Theta}\mathbf{\Psi}\mathbf{x}\mathbf{x}^H\mathbf{\Psi}^H\mathbf{\Theta}^H)$) has full rank.
- Once the parameters of the model have been estimated, the other set of the data is used to test the model. A mean square error is given. In general, a small mean square error is acknowledged to yield a good model and one seeks the smallest error possible.

If one is to use this procedure, one has to understand that in no way will it lead into judging the appropriateness of a model. Indeed, by adding more and more parameters to the model, one can always find a way of achieving a low mean square error by adjusting accordingly the parameters. This fact explains why some many models comply in the literature with the measurements. If the model minimizes the mean square error, then it is a **possible** candidate but the modeler can not conclude that it is a **good** candidate.

Moreover, since the testing method has no real justification, many problems arise when using it.

- How does one cut the set of data? Do we use half the data to estimate the parameters and half the data to test the model? Why not using one quarter and three quarter? In the Bayesian viewpoint, this is not at all a problem as one takes into account all the data available and does not make any unjustified transformation on the data.
- If one is to use a Music or Esprit algorithm, $\mathbf{K}$ has to be full rank. This is obviously not the case for a double directional model where the steering DoD matrix $\mathbf{\Psi}$ is not always full rank since $\mathbf{K} = \mathbb{E}(\mathbf{\Theta}\mathbf{\Psi}\mathbf{x}\mathbf{x}^H\mathbf{\Psi}^H\mathbf{\Theta}^H)$.

2- Moment fitting:

Other authors [47] validate their model by finding the smallest error of a set of moments. They derive explicit theoretical formulas of the n_{th} moment $m_n(f)$ of the matrix $\mathbf{H}^H(f)\mathbf{H}(f)$ and find the optimal parameters in order to minimize:

$$\frac{1}{N}\sum_{n=1}^{N} \mid \frac{m_n(f)}{\hat{m}_n(f)} - 1 \mid$$

where

$$\hat{m}_n(f) = \frac{\text{Trace}(\mathbf{H}^H(f)\mathbf{H}(f))^n}{\text{Trace}(\mathbf{H}^H(f)\mathbf{H}(f))}$$

As previously stated, many models can minimize this criteria by adding more and more parameters and one cannot obviously conclude in this case if a model is better then the other or not. Moreover, how useful is it to have a channel that fits a certain amount of moments?[26].

The previous remarks show that when the abstract of a paper asserts: "This paper finds the theoretical predictions to accurately match data obtained in a recent measurement campaign", one has to be really cautious on the conclusions to be drawn.

7 Conclusion

Where do we stand on channel modelling?[27] This question is not simple to answer as many models have been proposed and each of them validated by measurements. Channel models are not getting better and better but they only answer different questions based on different states of knowledge[28]. The crucial point is not

[26] Note that if all the moments fit, then the criteria is sound in the sense that measures such as mutual information or SINR (which are of interest in communications) will behave similarly.

[27] This question has to be taken in light of a talk "Where do we stand on maximum entropy?" made by E.T. Jaynes in 1978 at MIT [48].

[28] This point of view is not new and the misconception persists in many other fields. Descartes, already in 1637, warned us when stating in the first lines of the French essay "Le discours de la méthode":" la diversité de nos opinions ne vient pas de ce que les uns sont plus raisonnables que les autres, mais seulement de ce que nous conduisons nos pensées par diverses voies, et ne considérons pas les mêmes choses".

creating a model but asking the right question based on a given state of knowledge (raw measurement data, prior information, are we in a urban area? is it a fixed network?..). A generic method for creating models based on the principle of maximum entropy has been provided and proved to be theoretically sound. At every step, we create a model incorporating only our prior information and not more! The model achieved is broad as it complies as best it can with any case having more constraints (but at least includes the same prior constraints). The channel modelling method is summarized hereafter:

- $H(p) = \int -p\log p + \sum_i \lambda_i \{\text{prior information}\}_i$
- Argument of consistency

The consistency argument is extremely important as it shows that two channel modelling methods based on the same state of knowledge should lead to the same channel model. This fact has not always been fulfilled in the past. Our models are logical consequence of the use of the principle of maximum entropy and need not to be assumed without deeper justification. The models proposed may seem inadequate to reality for some readers: we argue as in [20] that the purpose of channel modelling is not to describe reality but only our information about reality. The model we achieve are consistent and any other representation is obviously unsound if based on the same state of knowledge. However, one must bear in mind that the less things are assumed as a priori information the greater are the chances that the model complies with any mismatched representation.

But what if the model fails to comply with measurements? The model is not to blame as it is a logic consequence of information theoretic tools [20]. With the methodology introduced, failure is greatly appreciated as it is a source of information and the maximum entropy approach is avid of information: the result of non-compliance is automatically taken into account as some new information evidence to be incorporated in the question. It only means that the question asked was not correct (double directional rather than directional for example) and should be adjusted accordingly in order to imply a new model (based on some new source of information); and as it is well known, finding the right question is almost finding the right answer.

References

1. G.J. Foschini and M.J. Gans, "On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas," *Wireless Personal Communications*, vol. 6, pp. 311–335, 1998.
2. K. Yu and B. Ottersten, "Models for MIMO Propagation Channels: A review," *Wireless Communications and Mobile Computing*, vol. 2, pp. 653– 666, November 2002.
3. H. Ozcelik, N. Czink, and E. Bonek, "What Makes a good MIMO Channel Model," in *Proceedings of the IEEE VTC conference*, 2005.
4. E. Biglieri, J. Proakis, and S. Shamai(Shitz), "Fading channels: Information-Theoretic and Communications Aspects," *IEEE Trans. on Information Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.
5. S. N Diggavi, N. Al-Dhahir, A. Stamoulis, and A. R. Calderbank, "Great Expectations: The value of Spatial Diversity in Wireless Networks," in *Proc. of the IEEE*, 219-270, Feb. 2004.
6. G. Golden, C. Foschini, R. Valenzuela, and P. Wolniansky, "Detection Algorithm and Initial Laboratory Results using V-BLAST Space-Time Communication Architecture," *Electronics Letters*, vol. 35, no. 1, pp. 14–16, Jan. 1999.
7. P.W. Wolniansky, G.J. Foschini, G.D. Golden, and R.A. Valenzuela, "V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel," in *International Symposium on Signals, Systems, and Electronics*, 1998, vol. 4, pp. 295–300.
8. I.E. Telatar, "Capacity of Multi-Antenna Gaussian Channels," Technical report, AT & T Bell Labs, 1995.
9. X. Giraud and J.C Belfiore, "Constellations Matched to the Rayleigh Fading Channel," *IEEE Trans. on Information Theory*, pp. 106–115, Jan. 1996.
10. J. Boutros and E. Viterbo, "Signal Space Diversity: a Power and Bandwidth Efficient Diversity Technique for the Rayleigh Fading Channel," *IEEE Trans. on Information Theory*, pp. 1453–1467, July 1998.
11. C. E. Shannon, "A Mathematical Theory of Communication," *The Bell Labs Technical Journal*, pp. 379–457,623–656, July-October, vol. 27 1948.
12. E. T. Jaynes, "Information Theory and Statistical Mechanics, Part 1," *Phys. Rev*, vol. 106, pp. 620–630, 1957.
13. E. T. Jaynes, "Information Theory and Statistical Mechanics, Part 2," *Phys. Rev*, vol. 108, pp. 171–190, 1957.
14. J. P. Burg, *Maximum Entropy Spectral Analysis*, Ph.D. thesis, Stanford University, 1975.
15. A. Zellner, *An Introduction to Bayesian Inference in Econometrics*, J. Wiley and Sons, New York, 2nd edition, 1971.
16. J. N Kapur, *Maximum Entropy Models in Science and Engineering*, John Wiley and Sons, Inc, New York, 1989.
17. G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Ph.D. thesis, Wahsington University, St. Louis, 1987.

18. H. Jeffrey, *Theory of Probability*, Oxford University Press, London, 1939, later editions, 1948, 1961.
19. J.M Keynes, *A Treatise on Probability*, MacMillan and Co., London, 1921.
20. E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge, 2003.
21. J.E Shore and R.W Johnson, "Axiomatic Derivation of the Principle of Maximum Entropy and The Principle of Minimum Cross-Entropy," *IEEE Trans. on Information Theory*, pp. 26–36, Jan. 1980.
22. R. T Cox, *Probability, Frequency and Reasonable Expectation*, Am. Jour. Phys., 14:1-13 edition, 1946.
23. Bretthorst G. Larry, "An Introduction to Model Selection Using Probability Theory as Logic," *in Maximum Entropy and Bayesian Methods, G. R. Heidbreder (ed), Kluwer Academic Publishers, Dordrecht the Netherlands*, pp. 1–42, 1996.
24. A. Mohammad-Djafari and G. Demoment, "Utilisation de l'Entropie dans les Problèmes de Restauration et de Reconstruction d'Images," *Traitement du Signal*, vol. 5, pp. 235–248, 1998.
25. M.A. Xapsos, G.P. Summers, and E.A. Burke, "Probability Model for Peak Fluxes of Solar Proton Events," *IEEE Transactions on Nuclear Science*, vol. 45, pp. 2948–2953, 1998.
26. H. Jeffreys, *Theory of Probability*, Oxford University Press, London, later editions, 1948, 1961 edition, 1939.
27. J. Boutros and G. Caire, "Iterative Multiuser Joint Decoding: Unified Framework and Asymptotic Analysis ," *IEEE Trans. on Information Theory*, pp. 1772–1793, July 2002.
28. T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 1991.
29. M. Franceschetti, S. Marano, and F. Palmieri, "The role of entropy in wave propagation," in *IEEE International Symposium on Information Theory*, Yokohama, Japan, July 2003.
30. M. Debbah M. Guillaud and A. L. Moustakas, "Analytical Channels," in preparation, 2006.
31. K. Liu, V. Raghavan, and A. M. Sayeed, "Capacity Scaling and Spectral Efficiency in Wideband Correlated MIMO Channels," *IEEE Trans. on Information Theory*, pp. 2504 – 2526, Oct. 2003 2003.
32. R. Müller, "A Random Matrix Model of Communication via Antenna Arrays," *IEEE Trans. on Information Theory*, pp. 2495–2506, Sep 2002.
33. R. Müller, "On the Accuracy of Modeling the Antenna Array Channel with Random Matrices ," in *International Symposium on Wireless Personal Multimedia Communications*, Aalborg, Denmark, 2001.
34. A. M. Sayeed, "Deconstructing Multiantenna Fading Channels," *IEEE Trans. on Signal Processing*, pp. 2563–2579, Oct. 2002.
35. J.P. Kermoal, L. Schumacher, K.I Pedersen, P.E. Mogensen, and F. Frederiken, "A Stochastic MIMO Radio Channel Model with Experimental Validation," *IEEE Journal on Selected Areas in Communications*, pp. 1211–1225, vol. 20, no. 6 2002.
36. D. Chizhik, J. Lingand P.W. Wolnianski, R. A. Valenzuela, N. Costa, and K. Huber, "Multiple-Input Multiple Output Measurements and Modeling in Manhattan," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 3 2002.
37. S.J. Fortune, D.H. Gay, B.W. Kernighan, O. Landron, R.A Valenzuela, and M. H. Wright, "WiSE Design of Indoor Wireless Systems: Practical Computation and Optimization," *IEEE Comput. Sci. Eng.*, vol. 2, pp. 58–68, Mar. 1995.
38. H. Ozcelik, M. Herdin, W. J. Weichselberg, and E. Bonek, "Deficiencies of "Kronecker" MIMO Radio Channel Model," *IEE Electronics Letters*, vol. 39, no. 16, pp. 1209–1210, Aug. 2003.
39. T.S. Pollock, "Correlation Modelling in MIMO Systems: When can we Kronecker?," *Australian Communications Theory Workshop*, vol. Newcastle, NSW, no. Australia, pp. 149–153, 2004.
40. D. Gesbert, H. Bölcskei, D. Gore, and A. Paulraj, "MIMO Wireless Channels: Capacity and Performance Prediction," *GLOBECOM conference records*, vol. 2, pp. 1083–1088, 2000.
41. D. Chizhik, G.J Foschini, M. J. Gans, and R. A. Valenzuela, "Keyholes, Correlations and Capacities of Multielement Transmit and Receive Antennas," *IEEE Trans. on Wireless Communications*, vol. 1, no. 2, pp. 361–368, Apr. 2002.
42. D. Gesbert, H. Bolcskei, D.A Gore, and A.J Paulraj, "Outdoor MIMO Wireless Channels: Models and Performance Prediction," *IEEE Trans. on Communications*, vol. 50, no. 12, pp. 1926–1934, Dec. 2002.
43. J. Capon, "High-Resolution Frequency Wavenumber Spectrum Analysis," *Proceedings of the IEEE*, vol. 8, no. 57, pp. 1408–1418, 1969.
44. R.O Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *Proceedings of RADC: Spectral Estimation Workshop, Rome*, pp. 243–258, 1979.
45. R. Kumaresan and D. W Trufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 134–139, 1983.
46. R. Roy and T. Kailath, "ESPRIT-Estimation of Signal Parameters via Rotational invariance techniques," *IEEE Transactions on Signal Processing*, vol. 7, pp. 984–995, 1989.
47. R. Müller and H. Hofstetter, "Confirmation of Random Matrix Model for the Antenna Array Channel by Indoor Measurements," in *IEEE Antennas and Propagation Society International Symposium, vol.1, pp.472-475*, Boston, Ma. USA, June 2001.
48. E. T. Jaynes, "Where Do We Stand on Maximum Entropy?," *in The Maximum Entropy Formalism,R. D. Levine and M. Tribus (eds.), M. I. T. Press, Cambridge, MA,*, p. 15, 1978.