An Information Geometric Approach to ML Estimation With Incomplete Data: Application to Semi-Blind MIMO Channel Identification

Amin Zia, James P. Reilly*, Jonathan Manton†, Shahram Shirani
Department of Electrical and Computer Engineering
McMaster University
Hamilton, Ontario, Canada L8S 4K1
Email: {ziama, reillyj, shirani}@mcmaster.ca
† Department of Information Engineering
Research School of Information Sciences and Engineering
The Australian National University, Canberra ACT 0200, Australia.
Email: jonathan.manton@rsise.anu.edu.au

Abstract

In this paper we cast the stochastic maximum likelihood estimation of parameters with incomplete data in an information geometric framework. In this vein we develop the *information geometric identification (IGID)* algorithm. The algorithm consists of iterative alternating projections on two sets of probability distributions (PD); i.e., likelihood PD's and data empirical distributions. A Gaussian assumption on the source distribution permits a closed form low–complexity solution for these projections. The method is applicable to a wide range of problems; however, in this paper the emphasis is on semi–blind identification of unknown parameters in a multi-input multi-output (MIMO) communications system. It is shown by simulations that the performance of the algorithm (in terms of both estimation error and bit-error-rate (BER)) is similar to that of the EM-based algorithm proposed previously [1], but with a substantial improvement in computational speed, especially for large constellations.

I. Introduction

With regard to identification of multi-input multi-output (MIMO) communication channels, if a training set consisting of input-output pairs to the channel is available, then the unknown parameters can be

∗Corresponding author

estimated using an ML estimation method that incorporates training. However, there are situations in which the observations do not include the input signals, and therefore the estimation must be carried out using only the available output observations. In such cases, since the observations alone are incomplete for estimating the unknown model parameters, the identification process, usually referred to as *blind identification*, relies on the available structure of the input as well as the signal model assumed. Blind identification problem in this case is based only on the partially available data, otherwise known as the *incomplete-data*. The EM algorithm [2] for solving the so-called *incomplete-data problem* is the main body of almost all algorithms that propose an approximate solution for stochastic blind identification. Previous work on the application of the EM algorithm in communications is presented in [1, 3–11].

In this paper we pose the incomplete data problem in an information geometric framework [12]. Information geometry encompasses a theoretical framework for a better understanding of estimation problems. The first paper that explicitly used the notion of information geometry for maximum likelihood estimation was due to Csiszar [13]. In this reference an iterative algorithm for minimizing the *Kullback-Liebler* (KL) distance between a given probability distribution representing the empirical distribution of the observations, and a family of probability distributions (the likelihood distributions) was proposed, and its relationship to ML estimation was investigated. Later in [14], an iterative algorithm for minimizing the KL-distance between two probability distribution (PD) convex sets was proposed and the application of the algorithm for maximum likelihood estimation was addressed. Maximum likelihood estimation with *incomplete-data* was posed as a double minimization of the KL distance between two PD sets in [15]. A similar approach was used for learning in the Boltzman machine [16] and for iterative image reconstruction [17]. The same problem was considered as a double minimization of the KL-distance between two sets of PD's in [18]. Specifically in these references, it was shown that this double projection information geometric approach is closely related to the EM algorithm [18, 19]. Interested readers are encouraged to refer to [16] for a review of the two approaches. In fact, the proposed IGID algorithm [20, 21] presented in this paper is shown to be identical to the *variational EM* algorithm [22].

The IGID algorithm uses a treatment similar to that given in [14]. The blind identification problem is implemented as a double minimization of the KL-distance between two PD sets. This minimization is realized in the form of iterative alternating projections. A major contribution of this paper is that closed-form solutions for these projections are developed. This closed-form nature of the algorithm is made possible by a Gaussian assumption on the source distribution[1]. The primary advantage of the proposed algorithm is computational. Previous EM algorithms for blind channel identification have not assumed Gaussianity of the source, and consequently suffer from computational complexity problems arising in the E-step, e.g., [1,4–8]. In contrast, the closed-form analytical projections used by the proposed IGID algorithm reduce computational costs significantly, especially for large constellations, with minimal degradation in performance. Thus, the proposed method inherits the asymptotically optimal properties of ML stochastic estimation, at substantially reduced cost. A previous closed–form EM algorithm which uses the Gaussian assumption, for blindly identifying single–input, single–output channels is presented in

[1]The impact of making such an assumption is discussed in Sect. IV.

[23].

Due to the similarity between information geometric alternating projection and EM algorithms, it would have been possible to derive a blind identification algorithm based on EM principles, that assumes a Gaussian source, instead of using information geometric principles. However, solving the blind identification problem in the information geometric framework gives new insight into the identification process and its relationship to the EM algorithm. Moreover, the development and the execution of the closed-form expressions for the required minimization (projection) operations are very straightforward and simple.

The organization of the paper is as follows. In Section II, the blind identification problem is reviewed, and the information theoretic interpretation of the problem in the form of a double minimization of the $KL-$distance between two PD sets is presented. The proposed method is applied to blind identification of a linear MIMO system with Gaussian noise in Section III where closed form solutions of the two alternating $KL-$distance minimizations are given. Simulation results are presented in Section IV. Section V includes some discussion on convergence and computational complexity issues, and Section VI concludes the paper.

Notation: Bold upper-case (lower-case) symbols indicate a matrix (vector) quantity respectively, while a symbol in calligraphic style indicates a set of probability distributions. The notation $N(\boldsymbol{\mu}, \boldsymbol{\Psi})$ denotes a multivariate (complex) normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Psi}$. In this paper, we consider joint distributions of the form $q(\boldsymbol{x}, \boldsymbol{y})$, and their associated marginal and conditional distributions $q(\boldsymbol{x})$ and $q(\boldsymbol{x}|\boldsymbol{y})$ respectively. Even though these are three distinct distributions, they are not denoted as such. The meaning of the distribution is evident from the structure of its argument. The subscript t is the IGID iteration index and k is the temporal index.

We refer to the following minimization

$$p^\star = \arg\min_{p\in\mathcal{P}} D(p||q) \tag{1}$$

as a type-I projection. It projects q onto $\mathcal{P}$, where q is an arbitrary PD, $\mathcal{P}$ is a set of PDs, and $D(p||q)$ is the $KL-$ distance measure, which is defined as

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{2}$$

where p and q in this case are discrete PDs and $q(x) \neq 0$ over the range of x. We refer to the following minimization

$$q^\star = \arg\min_{q\in\mathcal{Q}} D(p||q) \tag{3}$$

as a type-II projection, which projects p onto a PD set $\mathcal{Q}$. Because the KL distance measure is asymmetric, these two projections have different characteristics [24]. In the sequel, we often do not indicate the type of projection we are referring to. The type is made clear from the context.

II. Stochastic ML Estimation

Here we consider the general problem of *stochastic* maximum likelihood (ML) estimation of parameters. The stochastic ML approach assumes a class of PD for the unknown variables and therefore has the

advantage of imposing statistical structure on the inputs, in contrast to deterministic ML methods which ignore the statistics of the unknown quantities. For a comprehensive review of *deterministic* and *stochastic* methods to the blind identification problem, refer to [25, 26].

We consider the following linear time-invariant MIMO system with M transmitters and N receivers:

$$\boldsymbol{y}(k) = \sqrt{\frac{\rho}{M}}\boldsymbol{H}\boldsymbol{x}(k) + \boldsymbol{v}(k) \tag{4}$$

where $\boldsymbol{y}(k) \in \mathbb{C}^N$ and $\boldsymbol{x}(k) \in \Omega^M$ are the output and the input vectors, respectively, k is the time index, and Ω is a complex constellation with C members, such that the average energy over all members of the constellation is unity. The quantity $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ is the complex channel coefficient matrix, whose elements are zero mean random variables, scaled to unit *rms* values. The quantity ρ is the SNR on each receive channel. Also, the sources are chosen to be *i.i.d.*, whose components are zero–mean Gaussian. The quantity $\boldsymbol{v} \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$ is the noise vector with generally unknown covariance $\boldsymbol{\Psi} \in \mathbb{C}^{N\times N}$. It is assumed that $\boldsymbol{\Psi}$ is full rank.

The above MIMO model is valid in an intersymbol–interference (*ISI*) free Rayleigh–fading channel. It is assumed the channel $\boldsymbol{H}$ and the covariance $\boldsymbol{\Psi}$ are constant over a block of L transmitted symbols. This model is useful in space–time coding systems, where in many cases it is necessary to (semi) blindly identify the channel [27–29]. This model is also widely adopted in OFDM systems, e.g., [1].

A joint *pdf* of the input and output variables, e.g. $q(\boldsymbol{z};\boldsymbol{\theta})$, where $\boldsymbol{z} = [\boldsymbol{y}^T, \boldsymbol{x}^T]^T$ is the *complete* data, and $\boldsymbol{\theta} = (\boldsymbol{H}, \boldsymbol{\Psi})$ is the parameter set, provides a complete description of the underlying signal model. In general, for a given $\boldsymbol{z}$ there exists a one–to–one correspondence between $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $q(\cdot\,;\boldsymbol{\theta}) \in \mathcal{Q}$, where $\boldsymbol{\Theta}$ is the parameter space, and $\mathcal{Q}$ is the set of likelihood distributions, defined by

$$\mathcal{Q} = \{q(\boldsymbol{z};\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}. \tag{5}$$

The ML estimation task is then to choose a distribution in this family that best describes the complete data. By assuming that L independent complete data samples $\boldsymbol{z}_k,\ k = 1, \ldots, L$ are available, the maximum likelihood estimation problem is to find the distribution $q^*(\boldsymbol{z};\boldsymbol{\theta}^*)$ that satisfies

$$q^*(\boldsymbol{z};\boldsymbol{\theta}^*) = \arg\max_{q\in\mathcal{Q}} \prod_{k=1}^{L} q(\boldsymbol{z}_k;\boldsymbol{\theta}). \tag{6}$$

There are situations where the complete data are only partially available, i.e., we observe only $\boldsymbol{y}$. In these circumstances, the question is how to maximize the likelihood of observations and select the distribution $q(\boldsymbol{z};\boldsymbol{\theta}) \in \mathcal{Q}$ given only the partially available data (also called *incomplete–data*). Assuming that the input is discrete and distributed according to the *pdf* $p(\boldsymbol{x})$, one must solve the following equivalent *incomplete-data* problem:

$$q^*(\boldsymbol{y};\boldsymbol{\theta}^*) = \arg\max_{q\in\mathcal{Q}} \prod_{k=1}^{L} \sum_{\boldsymbol{x}_k} q(\boldsymbol{y}_k|\boldsymbol{x}_k;\boldsymbol{\theta})p(\boldsymbol{x}_k). \tag{7}$$

A. The Information Geometric Approach to Stochastic ML Estimation

We recall here how the ML estimate of (7) can be re–written in the form of a projection onto a set of distributions. The discrete distribution case is presented here for simplicity (extension to the continuous case is straightforward). Assume that the domain of incomplete observations $\boldsymbol{y}$ is divided into J neighborhoods, Δy_j; $j = 1, ..., J$ with $\tilde{\boldsymbol{y}}_j$'s as their center points. Now, given the observations $\boldsymbol{y}_k$, $k = 1, \ldots, L$, we define the *empirical distribution* $\tilde{p}$ of the observations to be:

$$\tilde{p}(\tilde{\boldsymbol{y}}_j) = \frac{1}{L} \sum_{k=1}^{L} \delta(\tilde{\boldsymbol{y}}_j - \boldsymbol{y}_k) \; ; \quad j = 1, ..., J \tag{8}$$

where $\delta(\cdot)$ is the Kronecker delta function defined as:

$$\delta(\tilde{\boldsymbol{y}}_j - \boldsymbol{y}_k) = \begin{cases} 1 \; ; & \boldsymbol{y}_k \in \Delta \boldsymbol{y}_j \\ 0 \; ; & \text{otherwise} \end{cases} \tag{9}$$

Using $q(\boldsymbol{y}_k) = \sum_{\boldsymbol{x}_k} q(\boldsymbol{y}_k, \boldsymbol{x}_k)$, the MLE problem (7) can be written as

$$\begin{aligned}
q^* &= \arg\max_{q \in \mathcal{Q}} \log(\prod_{k=1}^{L} q(\boldsymbol{y}_k)) \\
&= \arg\max_{q \in \mathcal{Q}} \sum_{k=1}^{L} \log q(\boldsymbol{y}_k) && (10) \\
&= \arg\max_{q \in \mathcal{Q}} \sum_{j=1}^{J} \sum_{k=1}^{L} \delta(\tilde{\boldsymbol{y}}_j - \boldsymbol{y}_k) \log q(\boldsymbol{y}_k) && (11) \\
&= \arg\max_{q \in \mathcal{Q}} L \sum_{j=1}^{J} \tilde{p}(\tilde{\boldsymbol{y}}_j) \log q(\tilde{\boldsymbol{y}}_j) && (12) \\
&= \arg\max_{q \in \mathcal{Q}} \left\{ \sum_{j=1}^{J} \tilde{p}(\tilde{\boldsymbol{y}}_j) \log \tilde{p}(\tilde{\boldsymbol{y}}_j) - \sum_{j=1}^{J} \tilde{p}(\tilde{\boldsymbol{y}}_j) \log \frac{\tilde{p}(\tilde{\boldsymbol{y}}_j)}{q(\tilde{\boldsymbol{y}}_j)} \right\} \\
&= \arg\max_{q \in \mathcal{Q}} \{-\mathcal{H}(\tilde{p}) - D(\tilde{p} \parallel q)\} && (13)
\end{aligned}$$

where $\mathcal{H}(\tilde{p})$ is the entropy of the empirical distribution $\tilde{p}$. Eq. (11) is obtained from (10) by the definition of the Kronecker delta function. Eq. (12) follows from (11) by using the definition of the empirical distribution given in (8). Since the entropy of $\tilde{p}$, i.e. $\mathcal{H}(\tilde{p})$, does not depend on the variable of maximization in (10) the ML estimation problem becomes:

$$q^* = \arg\min_{q \in \mathcal{Q}} D(\tilde{p}(\boldsymbol{y}) \parallel q(\boldsymbol{y})). \tag{14}$$

Thus, the ML estimation problem is equivalent to finding the projection of $\tilde{p}(\boldsymbol{y})$ onto the set $\mathcal{Q}$.

Observe that the optimization of (14) must find the best joint distribution $q(\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{Q}$ using only the information of the marginal distributions. To solve this incomplete data problem we proceed according to the method of [15] and define $\mathcal{P}$ as the set of all possible empirical distributions whose marginal

distribution over the unknown variable $\boldsymbol{x}$ is equal to the empirical distribution $\tilde{p}(\boldsymbol{y})$ of the observations:

$$\mathcal{P} = \{p(\boldsymbol{y},\boldsymbol{x}) | \sum_{\boldsymbol{x}} p(\boldsymbol{y},\boldsymbol{x}) = \tilde{p}(\boldsymbol{y})\}. \tag{15}$$

Now, for a given q_0, observe that:

$$\begin{aligned} D(p \parallel q_0) &= \sum_{\boldsymbol{y}} \sum_{\boldsymbol{x}} p(\boldsymbol{y},\boldsymbol{x}) \log \frac{p(\boldsymbol{y},\boldsymbol{x})}{q_0(\boldsymbol{y},\boldsymbol{x})} \\ &= \sum_{\boldsymbol{y}} \sum_{\left\{\boldsymbol{x} | \sum_{\boldsymbol{x}} p(\boldsymbol{y},\boldsymbol{x}) = \tilde{p}(\boldsymbol{y})\right\}} \tilde{p}(\boldsymbol{y}) p(\boldsymbol{x}|\boldsymbol{y}) \log \frac{\tilde{p}(\boldsymbol{y}) p(\boldsymbol{x}|\boldsymbol{y})}{q_0(\boldsymbol{y}) q_0(\boldsymbol{x}|\boldsymbol{y})}, \end{aligned} \tag{16}$$

where the last line follows because the joint distribution $p(\boldsymbol{y},\boldsymbol{x})$ representing the observed data and the corresponding inputs is physically constrained to lie within $\mathcal{P}$; hence $p(\boldsymbol{y},\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})\tilde{p}(\boldsymbol{y})$. The above derivation can be extended as follows:

$$\begin{aligned} D(p \parallel q_0) &= \sum_{\boldsymbol{y}} \tilde{p}(\boldsymbol{y}) \log \frac{\tilde{p}(\boldsymbol{y})}{q_0(\boldsymbol{y})} + \sum_{\boldsymbol{y}} \tilde{p}(\boldsymbol{y}) \sum_{\{\boldsymbol{x} | \sum_{\boldsymbol{x}} p(\boldsymbol{y},\boldsymbol{x}) = \tilde{p}(\boldsymbol{y})\}} p(\boldsymbol{x}|\boldsymbol{y}) \log \frac{p(\boldsymbol{x}|\boldsymbol{y})}{q_0(\boldsymbol{x}|\boldsymbol{y})} & (17) \\ &= D(\tilde{p}(\boldsymbol{y}) \parallel q_0(\boldsymbol{y})) + \mathrm{E}_{\tilde{p}(\boldsymbol{y})} D(p(\boldsymbol{x}|\boldsymbol{y}) \parallel q_0(\boldsymbol{x}|\boldsymbol{y})). & (18) \end{aligned}$$

Now, since q_0 and the empirical distribution of the observations $\tilde{p}(\boldsymbol{y})$ are given, the first term in (18) is unchanged by changing p. This term is thus a lower bound on the $KL-$distance between p and q_0. Therefore, the minimum $KL-$distance is achieved by letting $p(\boldsymbol{x}|\boldsymbol{y}) = q_0(\boldsymbol{x}|\boldsymbol{y})$ regardless of $\boldsymbol{y}$. This gives:

$$\min_{p \in \mathcal{P}} D(p \parallel q_0) = D(\tilde{p}(\boldsymbol{y}) \parallel q_0(\boldsymbol{y})) \quad \forall q_0 \in \mathcal{Q}. \tag{19}$$

Substitution of (19) in (14) gives:

$$\{q^*, p^*\} = \arg \min_{q \in \mathcal{Q}} \min_{p \in \mathcal{P}} D(p \parallel q). \tag{20}$$

Since the minimum $KL-$distance is achieved when $p(\boldsymbol{x}|\boldsymbol{y}) = q_0(\boldsymbol{x}|\boldsymbol{y})$, and by definition the marginal distribution of every distribution $p \in \mathcal{P}$ is equal to $\tilde{p}(\boldsymbol{y})$, one observes that $p^*(\boldsymbol{y},\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y})\tilde{p}(\boldsymbol{y}) = q_0(\boldsymbol{x}|\boldsymbol{y})\tilde{p}(\boldsymbol{y})$ achieves the minimum $KL-$distance in (19). This completes the proof for the following important theorem, which follows from [15]:

Theorem 1: Define the set $\mathcal{P}$ as in (15). Also define $\mathcal{Q}$ as the set of all likelihood PD's $q(\boldsymbol{y},\boldsymbol{x};\boldsymbol{\theta})$ each member of which is characterized by the parameter vector $\boldsymbol{\theta}$. The ML estimation of $\boldsymbol{\theta}$ can be obtained by the following double minimization:

$$\{q^*, p^*\} = \arg \min_{q \in \mathcal{Q}} \min_{p \in \mathcal{P}} D(p \parallel q). \tag{21}$$

Also, the projection of a given likelihood distribution $q_0(\boldsymbol{y},\boldsymbol{x})$ on $\mathcal{P}$ is given by:

$$p^*(\boldsymbol{y},\boldsymbol{x}) = \arg \min_{p \in \mathcal{P}} D(p \parallel q_0(\boldsymbol{y},\boldsymbol{x})) = q_0(\boldsymbol{x}|\boldsymbol{y})\tilde{p}(\boldsymbol{y}). \tag{22}$$

□

We therefore have the important result that stochastic ML estimation with incomplete data is equivalent to the double minimization of the $KL-$distance between the sets $\mathcal{Q}$ and $\mathcal{P}$. This double minimization is implemented using an iterative alternating projection method. After initializing with a suitable (p_0^*, q_0^*), at iteration t, $p_{t+1}^\star$ is the type-I projection of $q_t^\star$ onto $\mathcal{P}$, and $q_{t+1}^\star$ is the type-II projection of $p_{t+1}^\star$ onto $\mathcal{Q}$. Then, $t \leftarrow t+1$ and the process iterates until convergence.

If these two sets of PD's are convex[2], the double minimization has a global minimum, convergence to which is guaranteed [24]. However, in our case, $\mathcal{Q}$ is not convex and therefore the proposed alternating projection algorithm requires that the initial estimate $q_0^\star$ be determined through a proper initialization (training) procedure to assist convergence to the global minimum. The convergence analysis of the alternating projections approach to ML estimation of incomplete data is analogous to that of the EM algorithm. With the latter, conditions to guarantee convergence to a global optimum are very difficult to establish [30]. Thus, appropriate training sequences are used to alleviate convergence of the EM algorithm to the global optimum. The same applies to the IGID algorithm.

As can be seen in the following section, the projection onto $\mathcal{Q}$ is a convex optimization that gives a unique solution which is guaranteed to be a member of the set of desired likelihood distributions.

It is straightforward to observe that:

$$\begin{aligned} D(p_{t+1} \parallel q_{t+1}) &\leq D(p_{t+1} \parallel q_t) \\ &\leq D(p_t \parallel q_t), \end{aligned} \tag{23}$$

Since $D(\cdot \parallel \cdot)$ is bounded from below by 0, the sequence of pdf's generated by the algorithm decreases monotonically in $KL-$distance, and convergence to a local minimum is guaranteed.

III. Application to Semi–Blind Channel Identification

In this section, we apply the above information geometric approach to develop the computationally efficient IGID algorithm for semi–blind ML estimation of a multiple–input, multiple–output (MIMO) channel. The method is *semi-blind* due to the fact that the initial point is obtained by training the algorithm in each data block. We show the exact equivalence of the IGID algorithm and the *variational* EM algorithm [22] in Appendix I.

A. Signal Distributions

It is assumed that the input $\boldsymbol{x}(k) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Phi})$, where it is assumed $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\Phi}$ is assumed known. The set of likelihood distributions $\mathcal{Q}$ is parameterized by $\boldsymbol{\theta} = \{\boldsymbol{H}, \boldsymbol{\Psi}\}$, where $\boldsymbol{H}$ and $\boldsymbol{\Psi}$ are the channel, and the covariance matrix of the noise $\boldsymbol{v}$ in (4), respectively. Thus, each member of $\mathcal{Q}$ is a Gaussian likelihood distribution defined by:

$$q(\boldsymbol{z}; \boldsymbol{H}, \boldsymbol{\Psi}) = \mathcal{N}(\bar{\boldsymbol{z}}, \boldsymbol{Q}) \tag{24}$$

[2]For the definition of convexity, refer to [14].

where:

$$\bar{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{H}\boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \tag{25}$$

and

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{H}\boldsymbol{\Phi}\boldsymbol{H}^T + \boldsymbol{\Psi} & \boldsymbol{H}\boldsymbol{\Phi} \\ \boldsymbol{\Phi}\boldsymbol{H}^T & \boldsymbol{\Phi} \end{bmatrix} \tag{26}$$

where we have used the fact that $\boldsymbol{\Phi}^T = \boldsymbol{\Phi}$.

The expression for $\boldsymbol{Q}^{-1}$ is given for future convenience as (see Appendix (III):

$$\boldsymbol{Q}^{-1} = \begin{bmatrix} \boldsymbol{\Psi}^{-1} & -\boldsymbol{\Psi}^{-1}\boldsymbol{H} \\ -\boldsymbol{H}^T\boldsymbol{\Psi}^{-1} & \boldsymbol{\Phi}^{-1} + \boldsymbol{H}^T\boldsymbol{\Psi}^{-1}\boldsymbol{H} \end{bmatrix}. \tag{27}$$

For analytical and computational tractability, in the following we assume the empirical distribution corresponding to the observations is normally distributed; $\tilde{p}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{r}, \boldsymbol{S})$, where $\boldsymbol{r}$ and $\boldsymbol{S}$ are the mean vector and the sample covariance matrix for the output observations, respectively. This is in contrast to the exact form of empirical distribution given by (8). We note that since the source is assumed zero mean, and if we assume the channel has finite zero frequency gain, then under these assumptions we have $\boldsymbol{r} = \boldsymbol{0}$. This fact is used later.

B. The First Projection: Computing the Best Complete-Data Distribution

Due to the Gaussian source assumption, the complete data distribution is jointly Gaussian. Thus, this task is equivalent to finding the best mean and covariance of this joint distribution. Having the distribution $q_t(\boldsymbol{y}, \boldsymbol{x} : \boldsymbol{\theta})$ obtained from the previous iteration, we now solve the first minimization:

$$p^* = \min_{p \in \mathcal{P}} D(p \parallel q), \tag{28}$$

which is a type-I projection of the given PD q on the PD set $\mathcal{P}$.

The unique solution is given by the second part of Theorem 1. Therefore, to obtain the optimum distribution, one needs to compute the joint distribution of the complete-data likelihood distribution. In the case of jointly Gaussian PD's, straightforward mathematical manipulations yield the following closed form solution:

$$p^* = q(\boldsymbol{x}|\boldsymbol{y})\tilde{\boldsymbol{p}}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{P}^*), \tag{29}$$

where it can be shown that

$$\boldsymbol{m} = \bar{\boldsymbol{z}} + \boldsymbol{P}^*\overline{\boldsymbol{S}}^{-1} \begin{bmatrix} \boldsymbol{H}\boldsymbol{\mu} - \boldsymbol{r} \\ \boldsymbol{0} \end{bmatrix}, \tag{30}$$

$$(\boldsymbol{P}^*)^{-1} = \begin{bmatrix} \boldsymbol{\Psi}^{-1} - (\boldsymbol{H}\boldsymbol{\Phi}\boldsymbol{H}^T + \boldsymbol{\Psi})^{-1} + \boldsymbol{S}^{-1} & -\boldsymbol{\Psi}^{-1}\boldsymbol{H} \\ -\boldsymbol{H}^T\boldsymbol{\Psi}^{-1} & \boldsymbol{\Phi}^{-1} + \boldsymbol{H}^T\boldsymbol{\Psi}^{-1}\boldsymbol{H} \end{bmatrix} \tag{31}$$

and $\overline{\boldsymbol{S}}^{-1} \in \Re^{(M+N)\times(M+N)}$ is the covariance matrix $\boldsymbol{S}^{-1} \in \Re^{N\times N}$ properly augmented with zero blocks, i.e.:

$$\overline{\boldsymbol{S}}^{-1} = \begin{bmatrix} \boldsymbol{S}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \tag{32}$$

We note that, under the current assumptions, both $\boldsymbol{m}$ and $\bar{\boldsymbol{z}}$ are $\boldsymbol{0}$.

Therefore, to solve the first projection, i.e., to calculate $(\boldsymbol{P}^*)^{-1}$, it is sufficient only to modify the upper–left element of the inverse covariance matrix $\boldsymbol{Q}^{-1}$ given by (27), using the current estimates of $\boldsymbol{H}$ and $\boldsymbol{\Psi}$. The closed-form solution (31) avoids multi-dimensional integrations usually necessary in conventional EM-type algorithms. In Appendix I we show this projection is identical to the E-step of the EM algorithm.

C. The Second Projection: the Complete-Data ML Estimation

Given $p^*(\boldsymbol{z})$ from the previous projection, the second minimization is an ML estimation of the parameters, using the complete-data. The problem in the second minimization is to find the best distribution in the likelihood PD set $\mathcal{Q}$ that fits the estimated complete-data. This is equivalent to finding the type-II projection of p^* onto the PD set $\mathcal{Q}$. Therefore, it is necessary to solve the following minimization problem:

$$q^* = \arg\min_{q\in\mathcal{Q}} D(p^* \parallel q). \tag{33}$$

Since the $\mathcal{Q}$ family is parameterized by $\boldsymbol{H}$ and $\boldsymbol{\Psi}$, the optimization is performed with respect to these parameters. Assuming the following block form for the covariance matrix $\boldsymbol{P}^*$ of the given distribution p^*

$$\boldsymbol{P}^* = \begin{bmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} \\ \boldsymbol{P}_{12}{}^T & \boldsymbol{P}_{22} \end{bmatrix}, \tag{34}$$

the second projection is equivalent to the following minimization (see Appendix IV):

$$\begin{aligned} \{\boldsymbol{H}^*, \boldsymbol{\Psi}^*\} \;=\; & \arg\min_{\{\boldsymbol{H},\boldsymbol{\Psi}\}} \Big[\operatorname{trace}(\boldsymbol{\Psi}^{-1}\boldsymbol{P}_{11}) \\ & - \; 2\operatorname{trace}(\boldsymbol{\Psi}^{-1}\boldsymbol{H}\boldsymbol{P}_{12}^T) \\ & + \; \operatorname{trace}(\boldsymbol{\Phi}^{-1}\boldsymbol{P}_{22} + \boldsymbol{H}^T\boldsymbol{\Psi}^{-1}\boldsymbol{H}\boldsymbol{P}_{22}) \\ & - \; \log\det\boldsymbol{\Psi}^{-1} - \log\det\boldsymbol{\Phi}^{-1} \\ & - \; \log\det\boldsymbol{P}^* - d \Big], \end{aligned} \tag{35}$$

where $d = M + N$ is the dimension of the complete-data. Observe that by assuming a nonsingular noise covariance matrix $\boldsymbol{\Psi}$, the minimization is a convex optimization. It therefore has a unique solution. The minimization of the objective with respect to the parameters $\boldsymbol{H}$ and $\boldsymbol{\Psi}$ gives (see Appendix V):

$$\boldsymbol{H}^* = \boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1} \tag{36}$$

$$\boldsymbol{\Psi}^* = \boldsymbol{P}_{11} - \boldsymbol{P}_{12}\boldsymbol{P}_{22}^{-1}\boldsymbol{P}_{12}^T. \tag{37}$$

Therefore the iterative application of Equation (31), (36) and (37) generates the sequence of distributions $p_t, q_t, \quad t = 0, 1, \ldots,$ which in the limit yield maximum likelihood estimates for the model parameters $\boldsymbol{H}$ and $\boldsymbol{\Psi}$. It is shown in Appendix I that this projection is identical to the M-step of the EM algorithm.

D. Initialization Using Training

When a training data set consisting of input signal and output observation pairs are available, then identification is a *complete-data* ML estimation problem. Even though this problem is straightforward to solve in this case, it is interesting to note that it may be solved using an information geometric formulation. Maximum likelihood estimation corresponds to finding the closest (in the $KL-$distance sense) likelihood distribution $q(\boldsymbol{y}, \boldsymbol{x})$ to the empirical distribution of the input-output training data $\tilde{p}(\boldsymbol{y}, \boldsymbol{x})$:

$$q^* = \arg\min_{q \in \mathcal{Q}} D(\tilde{p} \parallel q) \tag{38}$$

Assume a set of L_{tr} training data pairs $\boldsymbol{z}_k = [\boldsymbol{y}_k, \boldsymbol{x}_k]^T, (k = 1, \ldots, L_{tr})$ is available. Assume that the empirical distribution (8) for the training data is modelled by a normal distribution, i.e. $\tilde{p}(\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{r}, \boldsymbol{S})$ where:

$$\boldsymbol{r} = \frac{1}{L_{tr}} \sum_{k=1}^{L_{tr}} \boldsymbol{z}_k \tag{39}$$

$$\boldsymbol{S} = \frac{1}{L_{tr}} \sum_{k=1}^{L_{tr}} \boldsymbol{z}_k \boldsymbol{z}_k^T = \frac{1}{L_{tr}} \sum_{k=1}^{L_{tr}} \begin{bmatrix} \boldsymbol{y}_k \boldsymbol{y}_k^T & \boldsymbol{y}_k \boldsymbol{x}_k^T \\ \boldsymbol{x}_k \boldsymbol{y}_k^T & \boldsymbol{x}_k \boldsymbol{x}_k^T \end{bmatrix}. \tag{40}$$

Then, it is straightforward to show that the estimates $\breve{\boldsymbol{H}}$ and $\breve{\boldsymbol{\Psi}}$ solving (38) are given by

$$\breve{\boldsymbol{H}} \sum_{k=1}^{L_{tr}} \boldsymbol{x}_k \boldsymbol{x}_k^T = \frac{1}{L_{tr}} \sum_{k=1}^{L_{tr}} \boldsymbol{y}_k \boldsymbol{x}_k^T \rightarrow \breve{\boldsymbol{H}} = \sum_{k=1}^{L_{tr}} \boldsymbol{y}_k \boldsymbol{x}_k^T \left(\boldsymbol{x}_k \boldsymbol{x}_k^T \right)^{-1}. \tag{41}$$

where $\boldsymbol{\Phi}$ is the source covariance matrix which is assumed known, and

$$\breve{\boldsymbol{\Psi}} = \frac{1}{L_{tr}} \sum_{k=1}^{L_{tr}} (\boldsymbol{y}_k - \breve{\boldsymbol{H}} \boldsymbol{x}_k)(\boldsymbol{y}_k - \breve{\boldsymbol{H}} \boldsymbol{x}_k)^T. \tag{42}$$

These results are similar to the least-squares solution for the ML estimation with training [31].

E. The IGID Algorithm: Summary

1. *Initialization*: Initial parameter estimates $\breve{\boldsymbol{H}}_0$ and $\breve{\boldsymbol{\Psi}}_0$ are obtained from the training data using (41) and (42) respectively.
2. *The first projection (onto $\mathcal{P}$), i.e., choosing the best empirical complete-data distribution:* In this step in iteration t, we compute the best distribution of the complete data $\boldsymbol{z} = [\boldsymbol{y}^T \boldsymbol{x}^T]^T$. Therefore we substitute the current values of $\boldsymbol{H}_t$ and $\boldsymbol{\Psi}_t$ into (31) to obtain the optimum covariance matrix $(\boldsymbol{P}^*)^{-1}$.
3. *The second projection (onto $\mathcal{Q}$), i.e., maximum likelihood estimation of parameters:* In this step, we compute the maximum likelihood estimate of the channel. Therefore, we invert $(\boldsymbol{P}^*)^{-1}$ and prepare the sub-blocks as in (34). Then we use Equations (36) and (37) to estimate the parameters $\boldsymbol{H}_{t+1}$ and $\boldsymbol{\Psi}_{t+1}$.
4. *Termination:* Check convergence by examining $D(p^* \parallel q^*)$ ((58)). If the distance is less than a predefined value $\epsilon, \quad 0 < \epsilon \ll 1$, terminate; otherwise, continue with the first projection (Step 2).

F. Convergence of the IGID Algorithm

We have seen that if the PD sets $\mathcal{P}$ and $\mathcal{Q}$ are convex, then the IGID algorithm converges [24]. However, in our case, these PD sets are Gaussian and hence are not convex. Even though (23) guarantees non-increasing divergence, it does not guarantee convergence to a single point in the intersection of $\mathcal{P}$ and $\mathcal{Q}$. As with the conventional EM algorithm, analysis of convergence of the IGID algorithm to a global optimum is difficult and is beyond the scope of this paper. Nevertheless, we can demonstrate the behaviour of the algorithm at convergence. In the following, we demonstrate there is a Gaussian distribution in $\mathcal{P}$ and one in $\mathcal{Q}$ that have equal means and covariances.

It is shown in Appendix VI that there exists a point $\{\hat{\boldsymbol{H}}, \hat{\boldsymbol{\Psi}}\}$ within $\mathcal{Q}$ such that

$$\boldsymbol{S} = \hat{\boldsymbol{H}}\boldsymbol{\Phi}\hat{\boldsymbol{H}}^{T} + \hat{\boldsymbol{\Psi}}, \tag{43}$$

where $\boldsymbol{S}$ is the covariance matrix of the observations $\boldsymbol{y}$ (defined in the paragraph under (27)). By substituting (43) into the upper–left block of (31) and substituting the values $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{\Psi}}$, it is straightforward to show that $(\boldsymbol{P}^{\star})^{-1}$ from (31) remains invariant from one iteration to the next. Further, $(\boldsymbol{P}^{\star})^{-1}$ from (31) is equal to $\boldsymbol{Q}^{-1}$ from (27), when $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{\Psi}}$ are substituted in (27). Thus, there exists a point in $\mathcal{P}$ and $\mathcal{Q}$ for which the covariance matrices of the distributions p_{∞} and q_{∞}[3] are equal and invariant with iteration.

From Section 3-B, we have seen that the mean $\bar{\boldsymbol{z}}$ of q_{∞} and the mean $\boldsymbol{m}$ of p_{∞} are both zero and invariant with iteration.

Therefore, since the means and covariances of the distributions are equal and are invariant with iteration, then under the stated conditions, there exists a point within $\mathcal{P}$ and $\mathcal{Q}$ to which convergence is possible. Thus, the convergence region of the IGID algorithm includes at least one point.

A consequence of this property is that $D(p_{\infty}||q_{\infty}) \rightarrow 0$.

IV. Simulations

A. Channel Estimation

In this section, we present simulation results for verifying the performance of the IGID algorithm for blind channel identification. The results are compared with a general previous EM-based algorithm which does not exploit a Gaussian source assumption, as summarized in Appendix II. ML estimation using all the data in the block as a training sequence is also performed, since this provides a lower bound on the performance of the algorithms. The IGID algorithm does not require the noise covariance to be known in general. However, in simulations we assume that the noise covariance matrix $\boldsymbol{\Psi}$ is known in order to be able to compare the results with previously reported algorithm. In each simulation, it is assumed that the channel gain matrix is constant within a block of length $L = 1000$ symbols.

In each block, the IGID and EM algorithms are initialized using training consisting of 10% of the block length, using (41) and (42). It is assumed that the symbols transmitted from each transmit antenna are selected uniformly from a 4-QAM (QPSK), 16-QAM or 64-QAM constellation, which has been normalized

[3]Recall the subscript on p or q refers to iteration index.

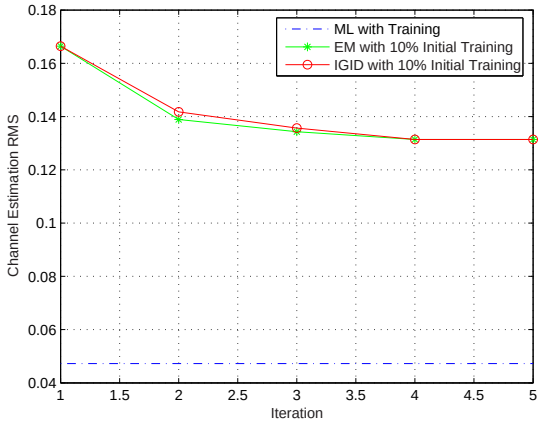

Fig. 1. Convergence (*rms* error vs. iteration index) for channel gain estimation in the low-SNR regime (SNR=6dB) in a 2 by 2 MIMO communication system with 16-QAM modulation. "ML with Training" uses the whole block of data as a training sequence, whereas the EM and the IGID algorithms each use 10% of the data block for training. The error is evaluated over 50 Monte Carlo runs.

to unit variance. The values M and N are each chosen to be equal to 2. The receiver noise covariance matrix $\boldsymbol{\Psi}$ is set to the identity $\boldsymbol{I}$. The channel coefficient matrix $\boldsymbol{H}$ is scaled corresponding to the desired values of SNR, as in (4). The channel coefficient matrix itself is chosen so that its elements are *iid* circular complex Gaussian random variables, with zero mean and unit variance.

The first index we use for quantifying the performance of the algorithms is the root mean–squared (*rms*) error of the channel estimate, at each iteration of the algorithm. This index shows the speed of convergence of the algorithms, as well as the extent to which the algorithm is able to converge to the true channel gain matrices. For these experiments, two values of SNR, arbitrarily chosen to be 16dB and 6dB, are chosen to demonstrate the performance of the algorithms in typical high and low-SNR conditions.

Figure 1 shows the *rms* error of the channel gain matrix estimation for 50 Monte-Carlo runs, for the IGID algorithm (circle), the EM algorithm (star), and ML estimation using the entire block as a training sequence (dashed line), versus iteration index, in the low-SNR regime ($SNR = 6$). The results are evaluated over $50,000$ symbols (50 blocks). It can be seen that the performance of the IGID algorithm is almost equal to that of the conventional EM algorithm. Also the IGID algorithm converges at a rate comparable to the EM-algorithm. Figure 2 shows the same results in the high-SNR regime ($SNR = 16$). Here, it can be seen that the IGID algorithm performance is only slightly degraded in terms of *rms* error compared to that of the EM algorithm.

B. Symbol-Error-Rate (SER)

To further examine the performance of the proposed algorithm, a symbol error rate (SER) analysis is performed. Each symbol is detected using an ideal ML procedure using the estimated values of $\boldsymbol{H}$ and $\boldsymbol{\Psi}$. Since in the simulations the number of sources is small, the usual complexity of ML detection is

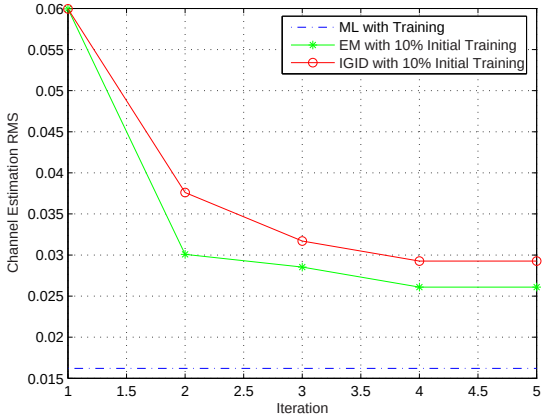

Fig. 2. Same as Figure 1, except SNR = 16 dB

.

tractable in this case. The estimated channel gain matrix and the noise covariance matrix are used for ML detection of the transmitted symbols $\boldsymbol{x}$, which are computed by:

$$\boldsymbol{x}^* = \max_{\boldsymbol{x}\in\Omega} p_x(\boldsymbol{x}|\boldsymbol{y}; \hat{\boldsymbol{H}}, \hat{\boldsymbol{\Psi}}) \tag{44}$$

where Ω is constellation set of the source. Figures 3–7 show SER results for 4QPSK, 16-QAM and 64-QAM modulation, each for block lengths of $L = 100$ and 1000 (only $L = 1000$ for the 64-QAM case). Each of these figures show SER results corresponding to the following estimation schemes for the parameters $\boldsymbol{H}$ and $\boldsymbol{\Psi}$: *i*) the parameters are perfectly known at the receiver (asterisk), *ii*) the parameters are estimated using only the initial training; i.e., EM and IGID are turned off (plus) *iii*) the parameters are estimated using the EM algorithm with 10% of the block used for training (x), and *iv*) the parameters are estimated using the IGID algorithm, again with 10% of the block of data used for training (circle). For the large L and small constellation case, it can be seen from the figures that the performance of the IGID algorithm is very close to that of the EM algorithm. However, it is seen that IGID's performance degrades slightly for decreasing values of L and increasing constellation size. For all the simulation scenarios considered in this study, the number of iterations for the IGID algorithm never exceeded five.

Simulations were performed where the percentage of training symbols relative to the block length was varied. It was determined that the SER performance is insensitive to training lengths above 10% when $L = 100$, and above roughly 5% when $L = 1000$. However, the number of iterations required for convergence changes somewhat with training length.

SER simulations were also conducted for the JADE [32] algorithm, which is deterministic due to the fact it does not exploit the specific distribution of the source nor the parameters. For the same simulation scenarios, it was found that the performance of the JADE method is up to about 3-4 dB worse than the IGID algorithm. This demonstrates the general idea that stochastic ML estimation is usually better than its deterministic counterpart.

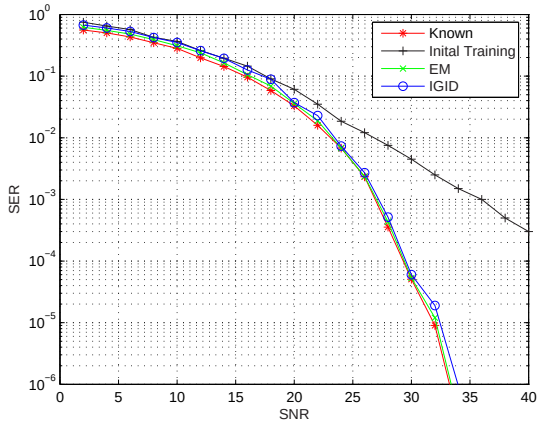

Fig. 3. Symbol Error Rate (SER) curves for QPSK modulation for a block length of $L = 100$.

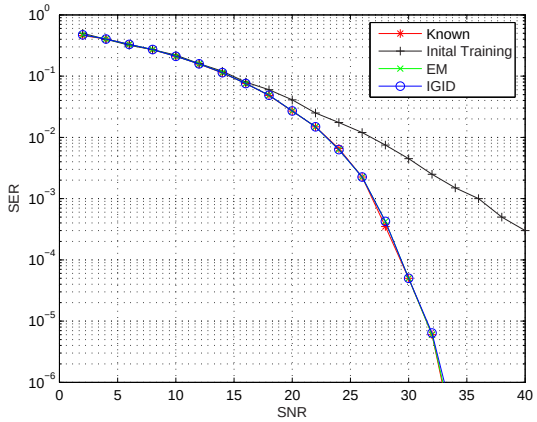

Fig. 4. Same as Figure 3, except the block length $L = 1000$.

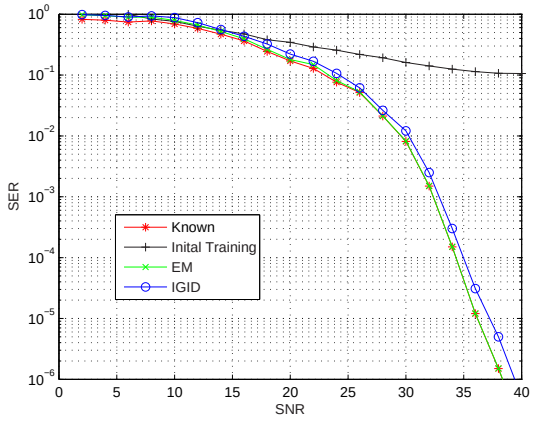

Fig. 5. SER curves for 16-QAM modulation for a block length of $L = 100$.

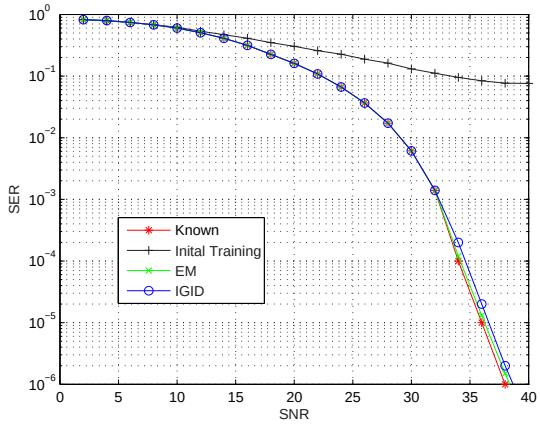

Fig. 6. SER curves for 16-QAM modulation for a block length of $L = 1000$.

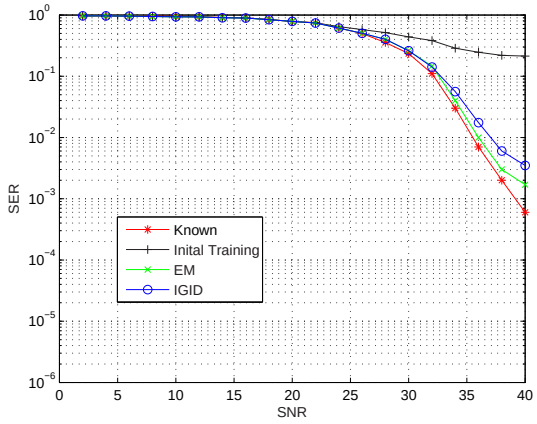

Fig. 7. SER curves for 64-QAM modulation for a block length of $L = 1000$.

In Figure 8, we show an example of the convergence of the IGID algorithm, for the same simulation scenario as in Figure 1. This figure shows the KL distance between p_t and q_t vs. the iteration index t. As expected, this measure converges monotonically towards zero.

C. Discussion

In addition to considering estimation performance, it is important to notice the considerable superiority in terms of computational complexity of the IGID algorithm over the EM-based algorithm. The complexity of the IGID algorithm is dominated by the inversion of the matrix $\boldsymbol{P}_{(M+N)\times(M+N)}$ in (31), which is on order of $\mathcal{O}((M+N)^3)$ using the Cholesky factorization [33], where M and N are the number of inputs and outputs, respectively. When the inputs are chosen from a set of discrete values, we see that the complexity of the IGID algorithm is independent of the number of input values and is on the order of $(M+N)^3$. This is in contrast to the complexity of the conventional EM-based algorithm, summarized in Appendix II, which is dominated by the computation of $\overline{\boldsymbol{x}\boldsymbol{x}^T}$ (defined in (54)), and is on the order of

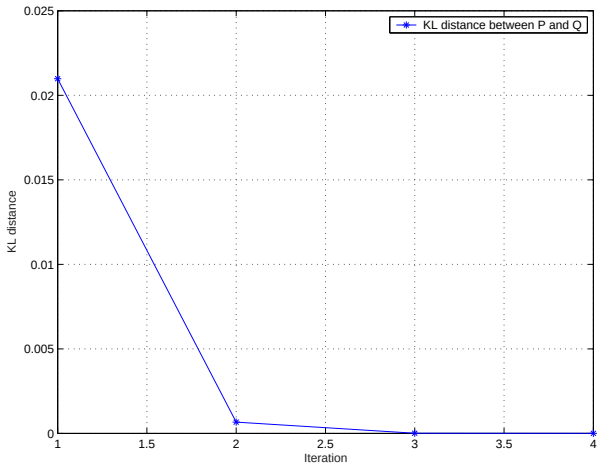

Fig. 8. $D(p_t||q_t)$ vs. iteration index t.

$\mathcal{O}(LM^2C^M)$ where L, M and C are the data block length, the number of input sources, and the number of discrete points in the constellation. Notice the exponential growth in complexity with C and M. Thus, for spectrally efficient signalling schemes, which use a large value of C, the IGID algorithm can be orders of magnitude faster than the EM algorithm.

For the current scenario using 16-QAM signalling, with $M = N = 2$, $C = 16$, and $L = 1000$, the complexity of the IGID algorithm is of the order 64, compared to $1,024,000$ for the EM-based algorithm. These figures are $(64; 51,200)$ and $(64; 10,240)$ when the block length reduces to $L = 500$ and $L = 100$ (appropriate for fast-fading channels), respectively. The ratio of the actual FLOP counts for each iteration of the two algorithms as measured by MATLAB© is about $(15,000)$, (800), and (150) for $L = 1000$, $L = 500$, and $L = 100$, respectively. This figures show a noticeable improvement in the execution speed of the proposed algorithm.

We have made the assumption that the source data and the observations are Gaussian distributed. This approximation is not far from reality in the low-SNR regime, due to the prominent effect of the noise on the output distribution. However, in high-SNR regime, since the noise effect is not as significant, the validity of this assumption diminishes. This seems to be the main reason for the discrepancy which exists between the performance of the IGID algorithm relative to that of previous EM algorithms. Nevertheless, this small deviation has a negligible effect on symbol error, as noted in the previous simulation results. However, we have noted that this marginal decrease in performance is accompanied by a significant gain in computational cost.

In this paper, the IGID algorithm has been developed for the following form of model:

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

where $\boldsymbol{y}(t)$ is a vector of observations, $\boldsymbol{H}$ is a matrix which is a function of unknown parameters, $\boldsymbol{x}(t)$ is the input vector, and $\boldsymbol{v}(t)$ is a noise vector. This model appears in many signal processing problems, such as direction of arrival estimation, tracking, model identification with hidden Markov models, in blind

identification of channels, etc. Thus, the proposed method is applicable not only to the blind identification problem, but to many other problems in signal processing as well.

V. Conclusions

In this paper an information geometric approach to blind identification was presented. Based on information geometry, a low-complexity iterative identification procedure, called the *IGID algorithm*, for blind identification of unknown parameters in a multi-input multi-output (MIMO) system with Gaussian distributed noise was proposed. The algorithm is an iterative solution to the *incomplete-data problem* posed by maximum likelihood (ML) estimation of parameters in a linear Gaussian MIMO system when only the output observations are available. The IGID algorithm involves two iterative minimizations, corresponding to projections onto the likelihood PD (*probability distribution*) set and the empirical PD set, respectively. A Gaussian assumption on the source allows us to develop closed–form expressions for the projection operations. The performance of the IGID algorithm in blind identification of the channel gain matrix in a MIMO communication system was investigated. Simulation results showing the symbol-error-rate (SER) behaviour are given. It is shown by simulation that the performance of the IGID algorithm is only slightly degraded relative to that of previous EM-based algorithms [1]; however, a noticeable improvement in computational cost is realized.

Appendix

I. Relationship Between the IGID and the Variational EM algorithm

In this Appendix we prove the equivalence of the IGID and EMS algorithms by introducing a variational version of the EM algorithm originally developed in [22]. Suppose that L input-output pair of observations $(x_1, y_1), ..., (x_L, y_L)$ are available. The standard maximum likelihood estimation aims to maximize the log-likelihood of the *complete-date* defined as:

$$\mathcal{L} = \log \prod_{k=1}^{L} q(y_k, x_k) = \sum_{k=1}^{L} \log q(y_k, x_k). \tag{45}$$

However, since the corresponding inputs of the observations are not available in the blind ML estimation problem, the procedure is based on only the output observations, which are also referred to as the incomplete-data:

$$\begin{aligned} \mathcal{L} &= \log \prod_{k=1}^{L} q(y_k) \\ &= \sum_{k=1}^{L} \log q(y_k) \\ &= \sum_{k=1}^{L} \log \sum_{x} q(y_k, x) \qquad (46) \\ &= \sum_{y} \tilde{p}(y) \log \sum_{x} q(y, x), \qquad (47) \end{aligned}$$

where the summation over x in (46) represents the set of all possible input values corresponding to the output observation. Also,(47) is obtained using the definition of the *empirical distribution* as defined in (8).

Using an arbitrary *variational* distribution over the input space, $u(x|y)$ we obtain a lower bound on the log-likelihood as follows:

$$\begin{aligned}
\mathcal{L} &= \sum_y \tilde{p}(y) \log \sum_x u(x|y) \frac{q(y,x)}{u(x|y)} && (48)\\
&\geq \sum_y \sum_x \tilde{p}(y) u(x|y) \log \frac{q(y,x)}{u(x|y)} && (49)\\
&= \sum_y \sum_x \tilde{p}(y) u(x|y) \log \frac{\tilde{p}(y) q(y,x)}{\tilde{p}(y) u(x|y)} \\
&= \sum_y \sum_x \tilde{p}(y) u(x|y) \log \frac{q(y,x)}{u(x|y)\tilde{p}(y)} + \sum_y \sum_x \tilde{p}(y) u(x|y) \log \tilde{p}(y) \\
&= \sum_y \sum_x \tilde{p}(y) u(x|y) \log \frac{q(y,x)}{u(x|y)\tilde{p}(y)} + \sum_y \tilde{p}(y) \log \tilde{p}(y) \\
&= \sum_y \sum_x p(y,x) \log \frac{q(y,x)}{p(y,x)} + \sum_y \tilde{p}(y) \log \tilde{p}(y) \\
&= -D(p \parallel q) - H(\tilde{p}(y)) \\
&\triangleq \mathcal{F}(p,q), && (50)
\end{aligned}$$

where $H(\tilde{p}(y))$ is the entropy of the observed empirical distribution. Eq. (48) is obtained from (49) using the *Jensen's Inequality* and the fact that the log function is convex.

The EM algorithm consists of two consecutive iterations [22] for maximizing the lower-bound $\mathcal{F}$ (50) as follows:

A. The E-Step vs. the First Projection

In the E-step, the best variational distribution over the input, $u(x|y)$ is computed to maximize the lower-bound (50):

$$\begin{aligned}
u(x|y) &= \arg\max_{u(x|y)} \mathcal{F}(p,q) \\
&= \arg\max_{u(x|y)} -D(p \parallel q) - H(\tilde{p}(y)) \\
&= \arg\min_p D(p \parallel q). && (51)
\end{aligned}$$

Eq. (51) follows from the fact that the entropy of the output empirical distribution $H(\tilde{p}(y))$ is constant.

Eq. (51) shows the similarity of the E-step to the *first projection* in the IGID algorithm. It is useful to observe that the solution of the first projection, i.e. $p^*(y,x) = q(x|y)\tilde{p}(y)$ achieves the equality in (50):

$$\begin{aligned}
\mathcal{F}(p^*, q) &= -D(p^* \parallel q(y,x)) - H(\tilde{p}(y)) \\
&= -D(q(x|y)\tilde{p}(y) \parallel q(y,x)) - H(\tilde{p}(y)) \\
&= -\sum_y \sum_x q(x|y)\tilde{p}(y) \log \frac{q(x|y)\tilde{p}(y)}{q(x|y)q(y)} + \sum_y \tilde{p}(y) \log \tilde{p}(y) \\
&= -\sum_y \tilde{p}(y) \log \frac{\tilde{p}(y)}{q(y)} + \sum_y \tilde{p}(y) \log \tilde{p}(y) \\
&= \sum_y \tilde{p}(y) \log q(y) = \mathcal{L}.
\end{aligned}$$

B. The M-Step vs. the Second Projection

The M-step of the EM algorithm maximizes the lower-bound (50) given the obtained distribution $u(x|y)$ in the E-step [22]. The maximization is performed with respect to the parameters of the likelihood function, which is equivalent to finding the best likelihood distribution q within the family of likelihood distribution $\mathcal{Q}$:

$$\begin{aligned}
q &= \arg\max_q \mathcal{F}(p, q) \\
&= \arg\max_q (-D(p \parallel q) - H(\tilde{p}(y))) \\
&= \arg\min_q D(p \parallel q).
\end{aligned}$$

This minimization corresponds to the *second projection* of the IGID algorithm.

II. EM-based Blind Identification Algorithms

EM methods have been developed for various blind identification problems [1, 3–9]. We summarize these the results as follows:

$$\boldsymbol{H}_{t+1} = \sum_{k=1}^{L} \overline{\boldsymbol{y}_k \boldsymbol{x}^T} (\sum_{k=1}^{L} \overline{\boldsymbol{x}\boldsymbol{x}^T})^{-1}. \tag{52}$$

$$\boldsymbol{\Psi}_{t+1} = \frac{1}{L} \sum_{k=1}^{L} \overline{(\boldsymbol{y}_k - \boldsymbol{H}_{t+1}\boldsymbol{x})(\boldsymbol{y}_k - \boldsymbol{H}_{t+1}\boldsymbol{x})^T}, \tag{53}$$

where $\boldsymbol{H}_{t+1}$ and $\boldsymbol{\Psi}_{t+1}$ are the new updated estimates of the channel gain matrix and noise covariance at iteration t and where:

$$\overline{\boldsymbol{x}\boldsymbol{x}^T} = \sum_{\boldsymbol{x} \in \Omega} \boldsymbol{x}\boldsymbol{x}^T p_x(\boldsymbol{x}|\boldsymbol{y}_k; \boldsymbol{H}_t, \boldsymbol{\Psi}_t) = \frac{\sum_{\boldsymbol{x} \in \Omega} \boldsymbol{x}\boldsymbol{x}^T p_y(\boldsymbol{y}_k|\boldsymbol{x}; \boldsymbol{H}_t, \boldsymbol{\Psi}_t)}{\sum_{\boldsymbol{x} \in \Omega} p_y(\boldsymbol{y}_k|\boldsymbol{x}; \boldsymbol{H}_t, \boldsymbol{\Psi}_t)}, \tag{54}$$

$$\overline{\boldsymbol{x}} = \frac{\sum_{\boldsymbol{x} \in \Omega} \boldsymbol{x} p_y(\boldsymbol{y}_k|\boldsymbol{x}; \boldsymbol{H}_t, \boldsymbol{\Psi}_t)}{\sum_{\boldsymbol{x} \in \Omega} p_y(\boldsymbol{y}_k|\boldsymbol{x}; \boldsymbol{H}_t, \boldsymbol{\Psi}_t)}, \tag{55}$$

and similarly

$$\overline{(\boldsymbol{y}_k - \boldsymbol{H}_{t+1}\boldsymbol{x})(\boldsymbol{y}_k - \boldsymbol{H}_{t+1}\boldsymbol{x})^T} = \frac{\sum_{\boldsymbol{x} \in \Omega} (\boldsymbol{y}_k - \boldsymbol{H}_{t+1}\boldsymbol{x})(\boldsymbol{y}_k - \boldsymbol{H}_{t+1}\boldsymbol{x})^T p_y(\boldsymbol{y}_k|\boldsymbol{x}; \boldsymbol{H}_t, \boldsymbol{\Psi}_t)}{\sum_{\boldsymbol{x} \in \Omega} p_y(\boldsymbol{y}_k|\boldsymbol{x}; \boldsymbol{H}_t, \boldsymbol{\Psi}_t)}, \tag{56}$$

where $\boldsymbol{H}_t$ and $\boldsymbol{\Psi}_t$ are the channel matrix and the noise covariance from the previous iteration, respectively. Also $p_y(\boldsymbol{y}_k|\boldsymbol{x};\boldsymbol{H}_t,\boldsymbol{\Psi}_t)$ is the likelihood function obtained by using the current values of the parameters $\boldsymbol{H}_t$ and $\boldsymbol{\Psi}_t$. These results are given assuming that the input consists of a finite set of points from constellation set Ω. Also since a uniform signalling scheme for the input is considered, the prior input distribution is $p_x(\boldsymbol{x}) = \frac{1}{C^M}$ where M is the length of input vector.

III. Proof of (27)

Defining the *Schur complement* of $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$ as $\boldsymbol{W} = Schur(\boldsymbol{M}) = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$, one can obtain the inverse of $\boldsymbol{M}$ by:

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{W}^{-1} & -\boldsymbol{W}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{W}^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{W}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix}. \tag{57}$$

Using the above, it is straightforward to derive the inverse of the block matrix $\boldsymbol{Q}$ in (27).

IV. Proof of (35)

It is easy to show that for Gaussian distributions $p = \mathcal{N}(0, \boldsymbol{P} \in \Re^{d\times d})$ and $q = \mathcal{N}(0, \boldsymbol{Q} \in \Re^{d\times d})$:

$$D(p \parallel q) = \mathrm{trace}(\boldsymbol{Q}^{-1}\boldsymbol{P}) + \log\det\boldsymbol{Q} - \log\det\boldsymbol{P} - d. \tag{58}$$

Substituting $\boldsymbol{Q}^{-1}$ and $\boldsymbol{P}$ from Equations (27) and (34), respectively, in (58) we have:

$$\begin{aligned} \mathrm{trace}(\boldsymbol{Q}^{-1}\boldsymbol{P}) &= \mathrm{trace}(\boldsymbol{\Psi}^{-1}\boldsymbol{P}_{11}) \\ &- 2\mathrm{trace}(\boldsymbol{\Psi}^{-1}\boldsymbol{H}\boldsymbol{P}_{12}^T) \\ &+ \mathrm{trace}(\boldsymbol{\Phi}^{-1} + \boldsymbol{H}^T\boldsymbol{\Psi}^{-1}\boldsymbol{H})\boldsymbol{P}_{22}. \end{aligned} \tag{59}$$

In addition since

$$\det\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \det\boldsymbol{A}.\det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}), \tag{60}$$

we have

$$\det\boldsymbol{Q}^{-1} = \det\boldsymbol{\Psi}^{-1}.\det\boldsymbol{\Phi}^{-1}. \tag{61}$$

Therefore:

$$\log\det\boldsymbol{Q} = -\log\det\boldsymbol{Q}^{-1} = -(\log\det\boldsymbol{\Psi}^{-1} + \log\det\boldsymbol{\Phi}^{-1}). \tag{62}$$

Substituting (58), (59) and (62) in (33) gives (35).

V. Proof of (36) and (37) [34]

From matrix algebra we have:

$$\frac{\partial(\mathrm{trace}(\boldsymbol{\Psi}^{-1}\boldsymbol{H}\boldsymbol{P}_{12}^T))}{\partial\boldsymbol{H}} = \boldsymbol{\Psi}^{-1}\boldsymbol{P}_{12} \tag{63}$$

$$\frac{\partial(\mathrm{trace}(\boldsymbol{H}^T\boldsymbol{\Psi}^{-1}\boldsymbol{H}\boldsymbol{P}_{22}))}{\partial \boldsymbol{H}} = 2\boldsymbol{\Psi}^{-1}\boldsymbol{H}\boldsymbol{P}_{22} \tag{64}$$

$$\frac{\partial(\mathrm{trace}(\boldsymbol{\Psi}^{-1}\boldsymbol{P}_{11}))}{\partial \boldsymbol{\Psi}^{-1}} = \boldsymbol{P}_{11}^T \tag{65}$$

$$\frac{\partial(\mathrm{trace}(\boldsymbol{\Psi}^{-1}\boldsymbol{H}\boldsymbol{P}_{12}^T))}{\partial \boldsymbol{\Psi}^{-1}} = \boldsymbol{P}_{12}\boldsymbol{H}^T \tag{66}$$

$$\frac{\partial(\mathrm{trace}(\boldsymbol{H}^T\boldsymbol{\Psi}^{-1}\boldsymbol{H}\boldsymbol{P}_{22}))}{\partial \boldsymbol{\Psi}^{-1}} = \boldsymbol{H}\boldsymbol{P}_{22}^T\boldsymbol{H}^T \tag{67}$$

$$\frac{\partial \log\det \boldsymbol{\Psi}^{-1}}{\partial \boldsymbol{\Psi}^{-1}} = \boldsymbol{\Psi}. \tag{68}$$

Using these equations to compute the partial derivatives necessary in the minimization of (35) results in (36) and (37).

VI. Proof of (43)

By definition, $\boldsymbol{P}$ is the covariance matrix of the complete data $\boldsymbol{z} = [\boldsymbol{y}^T\boldsymbol{x}^T]^T$. Therefore the true values of the blocks $\boldsymbol{P}_{22}$ and $\boldsymbol{P}_{11}$ in (34) are equal to $\boldsymbol{\Phi}$, the (true) known covariance matrix of the source $\boldsymbol{x}$, and $\boldsymbol{S}$, the sample covariance matrix of the observed data $\boldsymbol{y}$, respectively. From (36), we have

$$\boldsymbol{P}_{12} = \boldsymbol{H}^\star\boldsymbol{\Phi}.$$

Using this result in (37), we have

$$\begin{aligned}\boldsymbol{P}_{11} = \boldsymbol{S} &= \boldsymbol{\Psi}^\star + \boldsymbol{H}^\star\boldsymbol{\Phi}\boldsymbol{\Phi}^{-1}\boldsymbol{\Phi}^T\big(\boldsymbol{H}^T\big)^\star \\ &= \boldsymbol{H}^\star\boldsymbol{\Phi}\big(\boldsymbol{H}^T\big)^\star + \boldsymbol{\Psi}^\star.\end{aligned}$$

Therefore, there exist values $\hat{\boldsymbol{H}}$ and $\hat{\boldsymbol{\Psi}}$ in $\mathcal{Q}$ so that

$$\boldsymbol{S} = \hat{\boldsymbol{H}}\boldsymbol{\Phi}\hat{\boldsymbol{H}}^T + \hat{\boldsymbol{\Psi}}, \tag{69}$$

which was to be shown.

References

[1] C. H. Aldana, E. de Cardevalho, and J. Cioffi, "Channel estimation for miso systems using the em algorithm," *IEEE Transactions on Signal Processing*, vol. 51, pp. 3280–3292, 2003.

[2] N. M. L. A. P. Dempster and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[3] M. Feder and J. A. Capitovic, "Alorithm for joint channel estimation and data recovery- application tro equalization in underwater communications," *IEEE Transactions on Oceanic Engineering*, vol. 16, no. 1, pp. 42–55, 1991.

[4] X. Ma, H. Kobayashi, and S. C. Schwartz, "An enhanced channel estimation algorithm for OFDM: Combined EM algorithm and polynomial fitting," in *ICASSP 2003*, (Hong Kong), 2003.

[5] H. Zamiri-Jafarian and S. Pasupathy, "Recursive channel estimation for wireless communication via the em algorithm geometric derivation of em and generalized succesive interference cancellation algorithm,," in *ICPWC 1997*, 1997.

[6] R. A. Iltis and S. Kim, "Geometric derivation of em and generelizaed succesive interference cancellation algorithms with applications to cdma channel estimation," *IEEE Trans. Signal Processing*, vol. 51, pp. 1367–1377, May 2003.

[7] C. Cozzo and B. L. Hughes, "Joint channel estimation and data detection in space-time communications," *IEEE Trans. Communications*, vol. 51, pp. 1266–1270, Aug. 2003.

[8] L. Mazet, V. Buzenac-Settineri, M. de Courville, and P. Duhamel, "EM-based semi-blind estimation of time-varying channels," *IEEE Trans. on Signal Processing*, vol. 52, pp. 406–417, feb 2004.

[9] C. F. J. Wu, "On the convergence properties of the em algorithm," *The Annals of Statistics*, vol. 11, pp. 95–103, March 1983.

[10] A. Kocian and B. H. Fleury, "Em-based joint data detection and channel estimation of dc-cdma signals,," *IEEE Trans. on Communications*, vol. 51, pp. 1709–1720, Oct. 2003.

[11] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation maximization algorithm," *IEEE Trans. Signal Processing*, vol. 42, pp. 2664–2667, Oct. 1994.

[12] I. Csiszar, "Information theoretic method in probability and statistics," in *Available online at: http://ieeeits.org/publications/nltr/98_mar/01csi.pdf).*

[13] I. Csiszar, "$i-$divergence, geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.

[14] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures," *Statistics and Decision*, no. 1, pp. 205–237, 1984.

[15] W. J. Byrne, "Information geometry and maximum likelihood criteria," in *Conference on Information Sciences and Systems*, (Pfrinceton, NJ), 1996.

[16] W. J. Byrne, "Alternating minimization and Boltzman machine learning," *IEEE Trans. Neural Networks*, vol. 3, pp. 612–620, Apr 1992.

[17] C. Byrne, "Iterative image reconstruction algorithms based on cross-entropy minimization," *IEEE Trans. Image Processing*, vol. 2, pp. 96–103, 1993.

[18] S. I. Amari, "Information geometry of the EM and em algorithm for neural networks," *Neural Networks*, vol. 8, no. 9, pp. 1379–1408, 1995.

[19] S. I. Amari, K. Kurata, and H. Nagaoka, "Information geometry of the boltzman machine," *IEEE Trans. Neural Networks*, vol. 3, pp. 260–272, Mar 1992.

[20] A. Zia, J. P. Reilly, and S. Shirani, "Information geometric approach to channel identification: A comparison with em-mcmc algorithm," in *International Conference on Communications*, (Paris, France), 2004.

[21] A. Zia, J. P. Reilly, and S. Shirani, "An alternating projection method for blind joint channel identification and tracking," in *IEEE Statistical Signal Processing Workshop*, (St. Louis, Missouri), 2003.

[22] J. Roweis and Z. Ghahramani, *Learning Nonlinear Dynamical Systems using the Expectation-Maximization Algorithm, in S. Haykin, Ed., Kalman Filtering and Neural Networks*. New York: Dover Publications, 2001.

[23] J. H. Manton and Y. Hua, "Maximum-likelihood algorithms for deterministic and semi-blind channel identification," in *Second International Conference on Information, Communications and Signal Processing*, (Singapore), 1999.

[24] I. Csiszar and P. Shields, "Information theory and statistics: a tutorial," in *Available online at: http://www.math.utoledo.edu/ pshields/latex.html.*

[25] L. Tong and S. Perreau, "Multichannel blind identification: From subspace to maximum likelihood methods," *Proceedings of the IEEE*, vol. 86, pp. 1951–1968, Oct. 1998.

[26] E. de Cardevalho and D. T. M. Slock, "Blind and semi-blind fir multichannel estimation: (global) identifiability conditions," *IEEE TRans, Signal Processing*, vol. 52, pp. 1053–1063, 2004.

[27] S. M. Alamouti, "A simple transmiter diversity scheme for wireless communications," *IEEE J. Selected Areas in Communications*, vol. 16, pp. 1451–1458, Oct 1998.

[28] V. Tarokh, N. Seshardi, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criteria and code constructio," *IEEE Trans. Information Theory*, vol. 44, pp. 744–765, 1998.

[29] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *AT&T Bell Labs Internal Tech. Memo.*, 1995.

[30] C. F. J. Wu., "On the convergence properties of the em algorithm," *Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[31] J. K. Tugnait, L. Tong, and Z. Ding, "Single-user channel estimation and equalization," *IEEE Signal Processing Magazine*, vol. 17, pp. 17–28, May 2000.

[32] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE-Proceedings-F*, vol. 140, no. 6, pp. 362–370, 1993.

[33] G. H. Golub and C. F. V. Loan, *MATRIX Computations.* Baltimore, MA: The Johns Hopkins University Press, 2nd ed., 1993.

[34] K. B. Petersen and M. S. Pedersen, "Matrix cookbook," in *Available online at: http://matrixcookbook.com/*, vol. 6, 2006.