ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΟΙΚΟΝΟΜΙΑΣ,
ΑΝΑΠΤΥΞΗΣ & ΤΟΥΡΙΣΜΟΥ

ΕΝΙΑΙΟΣ ΔΙΟΙΚΗΤΙΚΟΣ ΤΟΜΕΑΣ
ΕΙΔΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΤΟΜΕΑΚΩΝ Ε.Π. ΤΟΥ ΕΚΤ

ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
Ε.Π. «ΑΝΑΠΤΥΞΗ ΑΝΘΡΩΠΙΝΟΥ ΔΥΝΑΜΙΚΟΥ,
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ»

Πίνακας με συνοδευτικές υποστηρικτικές πληροφορίες

Τίτλος πράξης	ΘΑΛΗΣ-ΕΚΠΑ-ΑΣΦΑΛΕΙΣ ΑΣΥΡΜΑΤΕΣ ΜΗ-ΓΡΑΜΜΙΚΕΣ ΕΠΙΚΟΙΝΩΝΙΕΣ ΣΤΟ ΦΥΣΙΚΟ ΕΠΙΠΕΔΟ
Κωδικός πράξης	380202
Τίτλος υποέργου	ΑΣΦΑΛΕΙΣ ΑΣΥΡΜΑΤΕΣ ΜΗ-ΓΡΑΜΜΙΚΕΣ ΕΠΙΚΟΙΝΩΝΙΕΣ ΣΤΟ ΦΥΣΙΚΟ ΕΠΙΠΕΔΟ
Τίτλος μελέτης (προσδορίστε από μελέτες, εκπαιδευτικό υλικό, εμπειρογνωμοσύνες, αξιολογήσεις κλπ)	D2.1 STATE-OF-THE-ART REPORT ON NONLINEAR REPRESENTATION OF SOURCES AND CHANNELS
Αριθμός συνημμένων αρχείων μελέτης	1
Τίτλοι υποπαραδοτέων μελέτης (σε περίπτωση που υπάρχουν)	
Ημερομηνία εκπόνησης της μελέτης	01/02/2012 - 30/04/2013
Τελικός δικαιούχος	ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
Φορέας υλοποίησης	ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
Ανάδοχος	ΚΑΘΗΓΗΤΗΣ ΝΙΚΟΛΑΟΣ ΚΑΛΟΥΠΤΣΙΔΗΣ

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

Επιχειρησιακό Πρόγραμμα
Ανάπτυξη Ανθρώπινου Δυναμικού,
Εκπαίδευση και Διά Βίου Μάθηση
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

MIS 380202

Secure **WI**reless **N**onlinear **COM**-munications *at the Physical Layer*

Deliverable 2.1

State–of–the–art report on nonlinear representation of sources and channels

S. Chouvardas, G. Giannakis, N. Kalouptsidis, I. Kontogiannis, A. Moustakas, E. Nistazakis, A. Stassinakis

European Union
European Social Fund

OPERATIONAL PROGRAMME
EDUCATION AND LIFELONG LEARNING
investing in knowledge society
MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS, CULTURE & SPORTS
MANAGING AUTHORITY
Co-financed by Greece and the European Union

NSRF 2007-2013
programme for development
EUROPEAN SOCIAL FUND

date:
March 11, 2014

version:
1.00

Contents

0 Preface

0.1 Scope and objectives of Work Package 2

The objective of Work Package 2 (WP2) is to establish a proper nonlinear modeling framework that will allow the development of novel tools and algorithms for the analysis and design of wireless communication systems. The expected outcomes within the context of this workpackage include (a) the development of novel techniques to estimate and compress sparse data sources through the use of parsimonious nonlinear models and (b) algorithms to estimate and sparsely represent nonlinear wireless channels. More generally, the framework developed in this workpackage sets the basic background for the remaining work packages and is expected to lead to: (a) Adaptive algorithms for nonlinear channel estimation and equalization. (b) Error probability bounds and appropriate coding schemes for the nonlinear wireless channel. (c) Cryptographic and encryption coding schemes of low-computational complexity, and enhanced security.

0.2 Report organization

This report consists of two complementary parts, related to the modeling of two important sources of nonlinearities in a communications system. In the first part, an overview of important past work related to the estimation, compression and processing of sparse data through the use of nonlinear models is provided. In the second part, the current state of the art on the representation of wireless channels in the presence of nonlinearities is summarized. In addition to the characteristics of the nonlinear wireless fading channel, some information is also provided on recent approaches to the sparse representation of such channels.

More specifically, Section 1.1 provides a general perspective on the need of models for sparse representation of signals, including speech and images, while Section 1.2 introduces the concepts of dimensionality reduction and sparse principal component analysis. Then, Section 1.3 briefly describes the main ideas behind compressive sampling and the introduction of adaptive algorithms like dictionary learning for the reconstruction of sparse data, and Section 1.4 provides an information-theoretic survey of issues related to the estimation and compression of discrete time series using variable-memory Markov chains and context tree weighting.

Subsequently, in Chapter 2, Section 2.1 briefly describes the important aspects of the wireless fading channel. Section 2.2 describes the current approaches to model channel nonlinearities, including the popular Volterra-type models and models for power clipping, while Section 2.4 extend these ideas to multi-antenna channels.

1 Representation and estimation of nonlinear sources

1.1 Source (De-) Coding: Context and Taxonomy

Source coding at the transmitter along with decoding at the receiver, aim at leveraging correlation among samples to remove redundancy, and thus store, process, compress, and transmit signals more efficiently, while being able to reconstruct them with minimal distortion at the receiver end. Given signal blocks (vectors), the first step of a source coding (a.k.a. compression or codex) module comprises a *dimensionality reduction* step of analog-amplitude vector samples. If transmitted via a digital communication system, this dimensionality reduction step is followed by a bit allocation (a.k.a. *quantization*) step, where a prescribed bit budget is optimally allocated (in some well defined sense) among the entries of the reduced dimensionality vector [5, 9, 17, 23].

Fundamental *information-theoretic limits* for lossy compression and reconstruction are offered by the celebrated rate-distortion theory, and its generalizations to remote sources and distributed settings dealing with the so-termed CEO problem, which entails (de-)coding with side information [36, 47, 49, 46, 14]. Given a number of bits and a reconstruction metric (e.g., mean-square error), rate-distortion theory asserts that for Gaussian distributed signals the process of Karhunen-Loeve based dimensionality reduction followed by the reverse waterfilling allocation of bits attains the lowest error for the prescribed rate [5, 17]. On the practical side, mpeg and jpeg standards implement state-of-the-art (de)coders, which rely on the Fourier or the wavelet transform, and iterative (trellis) scalar or vector quantization.

Low-dimensional *modeling* and dimensionality reduction are cornerstone tasks in various areas, where high-dimensional vectors must be mapped to their judiciously chosen low-dimensional counterparts – what is also referred to as space (or manifold) learning and embedding [27]. In addition to compression, model reduction, system identification, pattern recognition, big data processing and storage, as well as machine learning tasks (including regression, classification, feature extraction, and clustering), all continuously benefit from advances in low-dimensional modeling and dimensionality reduction [24]. Many of these tasks operate only on the low-dimensional (sub)space(s), while (de)coders must eventually bring compressed vectors back to their high-dimensional reconstructed counterparts.

Over the last decade, researchers recognized two possibilities that spurred renewed interest in source (de)coding. First, the fact that sampled signals can be sparse (and thus entail low-dimensional structure) when expanded over an appropriate basis. This attribute of natural or man-made *sparsity* can be exploited fruitfully in compression and reconstruction [16, 24]. Second, it was appreciated that source coding of sparse signals could be pursued jointly with sampling of continuous-variable signals, what led to the popular themes of *compressive sampling* or compressed sensing (CS) [10, 18, 3], and the more recent efforts on sparse coding via *dictionary learning* (DL) [35, 42]. The emergent tools can be also utilized for robust source (de)coding when signals are observed in the presence of outliers, and also when samples are missing. These CS and DL advances to joint sampling and compression are intimately related to the basic problems of interpolation (a.k.a. imputation), extrapolation (a.k.a. prediction) themes, but they are also at the heart of contemporary subjects, including matrix completion [11], low-rank representation [13], and their popular applications to recommender systems and webpage rank schemes [24].

In a nutshell, the vast majority of prior (de-)coding works has been confined to *linear* source models, *linear* dimensionality reducing operators to obtain low-dimensional embeddings, and *linear* reconstruction operators at the receiver end. Well justified for Gaussian stationary vector processes, these operators are obtained based on the sample covariance matrix found using training vectors, and in this sense they are *data-adaptive*. The latter is to be contrasted with the CS approaches, which rely on a so-termed measurement matrix that is chosen to satisfy special conditions (known as restricted isometry properties), but otherwise it is *data-nonadaptive* since it ignores the underlying statistics of the signals to be compressed. There is clearly a need for source (de-)coding approaches suitable for non-Gaussian, and (at least piecewise) nonstationary signal sources, which are capable of coping with low-dimensional *nonlinear* models and embeddings in a *data-adaptive* manner, but also offer computationally affordable compression, reconstruction, and learning algorithms.

1.2 Low-dimensional modeling and dimensionality reduction

Consider $N \in \mathbb{N}_*$ high-dimensional (column) vectors $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{R}^D$, located on or close to a smooth but otherwise unknown manifold $\mathcal{M} \subset \mathbb{R}^D$, $D \in \mathbb{N}_*$. Given these training data vectors, critical for efficient source encoding and decoding of out-of-sample vectors $\boldsymbol{x}$ are: (a) the dimensionality reduction module, which effects (generally lossy) compression from high-dimensional ($\boldsymbol{x} \in \mathbb{R}^D$) to low-dimensional ($\boldsymbol{y} \in \mathbb{R}^d$, $\mathbb{N}_* \ni d \ll D$) vectors at the transmitter (Tx); as well as (b) the reconstruction module at the receiver (Rx), which yields estimates $\hat{\boldsymbol{x}}$ of the high-dimensional vectors from their low-dimensional renditions.

Principal component analysis (PCA) relies on the Karhunen-Loeve transform, which constitutes the "workhorse" of dimensionality reduction using a *linear* operator, namely a $d \times D$ matrix $\boldsymbol{U}^T$ (^{T} denotes transposition) formed by the eigenvectors corresponding to the d (out of D) largest eigenvalues of the sample covariance matrix $N^{-1} \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^T$ [24, Chap. 14.5]. As the latter is formed using training data, PCA is a data-adaptive operation. PCA's premise for compressing $\boldsymbol{x}$ to its lower-dimensional rendition $\boldsymbol{y} = \boldsymbol{U}^T \boldsymbol{x}$ at the Tx, and reconstructing it optimally, in the mean-square sense, is that $\boldsymbol{x}$ is stationary with the same covariance matrix as $\{\boldsymbol{x}_n\}_{n=1}^N$. From a deterministic viewpoint, PCA is effective in (de)compression provided that both training and out-of-sample vectors live on (or stay close in the least-squares (LS) sense to) an affine subspace.

Albeit never explicitly used, the low-dimensional model underlying PCA is the following linear one: $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{y} + \boldsymbol{e}$, where $\boldsymbol{U}$ has size $D \times d$ and $\boldsymbol{e}$ denotes white noise. Over the last dozen years, a surge of research has emerged that exploits the attribute of sparsity, which in the present context presumes that $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{s}$, where $\boldsymbol{s}$ is a sparse vector (a number of its entries are zero but their locations are unknown) over a certain basis $\boldsymbol{B}$. The first cluster of past works on this topic deals with sparse PCA; see e.g., [27]. However, the nonconvex criterion involved in these works does not lend itself to efficient optimization. Improved optimization algorithms are reported in [52] using block coordinate descent [6]; and also in [2] using relaxation and greedy solvers. Related approaches augment the standard singular value decomposition (SVD) cost, or, the maximum likelihood criterion with ℓ_1 norm penalties to effect sparsity [44, 45, 12]. All the aforementioned sparse PCA schemes neither exploit sparsity for compression and reconstruction nor they account for non-ideal encoder-to-decoder links, and power constraints at the encoder side. Preliminary results on the latter can be found in [38].

An important generalization of low-dimensional models is offered by those capturing "union of subspaces" and "nonparametric matrix factor analysers," which can be estimated using either maximum likelihood or Bayesian techniques; see e.g., [15, 13] and references therein. The motivation behind these nonlinear models is that data vectors do not generally lie on an affine subspace, but often on a manifold. In addition, they are typically realizations of nonstationary or locally stationary processes, including those formed by e.g., image and speech signals. These considerations prompt approaches to *nonlinear* dimensionality reduction – a subject explored over the last dozen years primarily in the context of machine learning themes such as clustering and feature extraction [24, Chap. 14.9]. Among those, popular ones are the multidimensional scaling (MDS), the ISOMAP, the locally linear embedding (LLE), the Laplacian Eigenmaps (LLE), the semidefinite embedding (SDE), and their common spectral decomposition tool known as kernel PCA [39, 22, 24]. Among these, LLE has well-documented merits, because [37]: (a) it is computationally affordable, entailing closed-form expressions and eigen-decomposition level complexity; (b) it does not require knowledge but only smoothness of the manifold; and (c) it leverages smoothness to learn the manifold, and obtain LLEs that can be thought of as being applied on tangential affine subspaces.

So far, LLE has been advocated for manifold learning, clustering, and classification [22]. Recently, sparsity has been leveraged to render (kernel) PCA and MDS robust against outliers [33, 20], but also for LLE-type robust manifold learning and low-dimensional embedding [25, 41, 19, 28]. Neither one has been investigated for reconstruction purposes in a source (de)coding setup. Sparsity is also the enabling attribute for compressive sampling (CS) via random projections and also for dictionary learning (DL), two subjects outlined next.

1.3 Compressive sampling and dictionary learning

Sparsity is an attribute characterizing many natural and man-made signals, not only because nature is inherently parsimonious, but also because practical constraints encourage engineering designs with as few degrees of freedom as possible. For this reason, sparsity has been exploited over the last dozen years in a broad range of *statistical* inference tasks concerned with the choice of most informative variables in linear regression models using e.g., the least-absolute shrinkage and selection operator (Lasso) [24]. Lasso has been successfully adopted in a gamut of applications, ranging from the discovery of behavioral trends in social networks, to unveiling interpretable biological structure in gene expression micro-array data, diabetes, and prostate cancer prognosis [24]. In parallel, related basis pursuit ideas have capitalized on sparsity to obtain parsimonious signal representations primarily in *deterministic* settings [16]. The latter have led to the recent ground-breaking results on CS, where sparsity has been proved instrumental to solving under-determined *linear* systems of equations. CS has created excitement in signal processing circles too, for sub-Nyquist sampling of sparse signals [10]. Experimental demonstrations of CS have also emerged using *analog* designs [3]. To deal with disturbances and noise, Lasso, basis pursuit, and CS algorithms all involve minimization of a squared-error cost regularized by the ℓ_1 norm of the unknowns.

Sparse PCA approaches (and natural extensions to sparse canonical correlation analysis (CCA)) exploit sparsity present in second-order statistical descriptors of the stationary processes involved. Dimensionality reduction based on them will yield high reconstruction performance provided that the (cross-) covariance or spectral density matrices of the data are sparse, and also representable by a few strong eigen-components. However, there are cases where the data themselves admit a sparse *deterministic* representation over a perhaps unknown basis that is not necessarily generated by the eigenspace of the covariance or the power spectral density matrix. For instance, images tend to admit sparse representations over the wavelet basis [1]. If such a basis were known, then it could be used to represent the data with a coefficient vector whose dimensionality is smaller than the one of the original data vector [16]. Using training data to learn the underlying basis, recent works have advocated sparse overcomplete basis expansion schemes to parsimoniously represent and effectively reconstruct data vectors [35, 1, 32]. The resultant so-termed dictionary learning (DL) algorithms are data-adaptive and require solving a sparsity-aware, bilinear regression problem. The proposed research will capitalize on deterministic descriptors of sparsity to develop DL-type approaches that account for power constraints at the encoder, and non-ideal encoder-to-decoder links. It will also accommodate data vectors admitting sparse representations over time-varying overcomplete bases – a case of paramount importance for compression of nonstationary (e.g., video) sources.

Summarizing, sparsity is expected to play an instrumental role in various inference tasks. It has been so far exploited for CS [4] [48] [50], DL [35, 42], and reconstruction. However, its role for LLE-like nonlinear, data-adaptive (de)compression has not been investigated, except for our preliminary work reported in [40]. Performance analysis is a fertile ground for research, and fundamental (even asymptotic) limits are yet to be established.

1.4 Information theoretic issues and compression

In this section, our focus will be on the analysis, estimation and compression of nonlinear sources, based primarily on *sparse representations*. Moreover, the main focus of the present section – as well as the focus of the relevant proposed work – will be on *discrete* sources, since the discreteness of the signal is often the main reason for its “nonlinearity.”

Before proceeding to review the existing relevant literature in detail, we briefly mention that there are connections with the well-established and extensively developed area of *vector quantization* of continuous sources. Until the late 1990s, most of the research effort in vector quantization was devoted to addressing the issue of universality, see [76] and the references therein, as well as [77] [78] [79] [80][81] [82] [83][84]. Algorithms emphasizing more practical aspects have been proposed in [85][86][87][88]. So-called “structured source codes” and “structured vector quantizers” also offer sparse representations. In this connection, we mention tree codes developed by Jelinek and others [89] [90] [85, Ch. 15]; and source codes based on trellises [91] [92] [93] [94]. Also, there is a long line of work on compression algorithms based on linear codes, of which the most complexity-efficient ones are those that combine a linear code with an encoder utilizing sparse-graph properties or a message-passing-type algorithm; see, e.g., [95][96] and the references therein. Finally, we mention that, more recently, a new class of codes using sparse random dictionaries have been proposed in [97][98], extending corresponding channel coding schemes [99], all of which are based on sparse regression ideas from statistics.

1.4.1 Estimation of sparse discrete sources

The main problem of modeling and compression of discrete time series is the fact that, long memory – the most important feature of the data which can be utilized for effective compression – cannot be modeled effectively. The obvious description of a model as a dth order Markov chain has long been recognized as problematic: For example, in order to describe the model of Markov data with an alphabet of size m, with memory length d, requires the estimation of at least $(m-1)m^d$ parameters; even with very moderate values for m and d this is obviously prohibitively large. For example, a Markov chain with memory length $d = 20$ with an alphabet of size $m = 10$, requires the estimation of $\approx 10^{21}$ parameters, clearly an outrageously impractical goal. See, for example, the extensive discussions in [100][101] and the references therein.

The most successful line of research that has dealt with this high-dimensionality problem is that of the utilization of so-called *tree sources* or *variable-memory Markov chains*. We first briefly review their structure and then we give bibliographical pointers to where relevant theoretical results as well as applications have been developed.

Our starting point is the definition of the distribution of class of dth order, homogeneous Markov chains, with values in the finite state-space, or “alphabet” $A = \{0, 1, \dots, m-1\}$. The memory length $d \geq 0$ and the alphabet size $m \geq 2$ are fixed.

Specifically, for the process $\{X_n\}$ $H = \{X_{-D}, X_{-D+1}, \dots, X_{-1}, X_0, X_1, \dots\}$, we will define the conditional distribution of each X_i, $i \geq 1$, given the previous d symbols $(X_{i-d}, X_{i-d+1}, \dots, X_{i-1})$, where we write X_i^j for a vector of random variables $(X_i, X_{i+1}, \dots, X_j)$ and similarly $x_i^j \in A^{j-i+1}$ for a string $(x_i, x_{i+1}, \dots, x_j)$ representing a realization of the random variables X_i^j. The key element in specifying these distributions is a *context function* $C{:}A^d \to T$, which maps each length-d context x_{i-d}^{i-1} to a (typically strictly) shorter suffix $C(x_{i-d}^{i-1}) = x_{i-j}^{i-1}$ of itself, for some $0 \leq j \leq d$. Then the Markov property for $\{X_n\}$ takes the form:

$$P(x_1^n | x_{-d+1}^0) = \prod_{i=1}^{n} P(x_i | x_{i-d}^{i-1}) = \prod_{i=1}^{n} P(x_i | C(x_{i-d}^{i-1})).$$

The range T of C is a subset of $\cup_{d=0}^{d} A^d$, where we adopt the convention that the set A^0 contains only the empty string λ. We assume that the set T is *proper*, namely, that no element in T is a proper suffix of any other, and that if some $x_i^j = (x_i, x_{i+1}, \dots, x_j)$ is in the range of C, then so is every string of the form $(y, x_{i+1}, \dots, x_j)$, for all $y \in A$. Observe that, under these assumptions, the context function C is completely determined by its range T, since, for any string x_{i-d}^{i-1} there is exactly one element of T which is a suffix x_{i-j}^{i-1} of x_{i-d}^{i-1}.

To complete the specification of the (conditional) distribution of the process $\{X_n\}$, in addition to the *context set* T, with every element $s \in T$ we associate a probability vector $\theta_s = (\theta_s(0), \theta_s(1), \dots, \theta_s(m-1))$, [where the $\theta_s(j)$ are nonnegative and sum to one, $\sum_{j\in A}\theta_s(j) = 1$]. Then, the probability $P(x_1^n|x_{-d+1}^0)$ is,

$$P(x_1^n|x_{-d+1}^0) = \prod_{i=1}^n P(x_i|x_{i-d}^{i-1}) = \prod_{i=1}^n P(x_i|C(x_{i-d}^{i-1})) = \prod_{i=1}^n \theta_{C(x_{i-d}^{i-1})}(x_i). \tag{1.1}$$

Note that, instead of taking the product sequentially in time, we can instead take a product over all possible contexts $s \in T$, and express this probability as,

$$P(x_1^n|x_{-d+1}^0) = \prod_{s\in T}\prod_{j\in A}\theta_s(j)^{a_s(j)}, \tag{1.2}$$

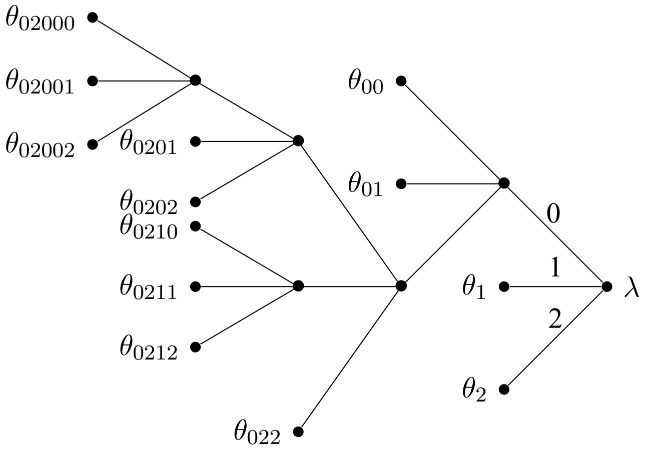
where each element $a_s(j)$ of the vector $a_s = (a_s(0), a_s(1), \dots, a_s(m-1))$ is,

$$a_s \;=\; \#\text{ times symbol } j \in A \text{ follows context } s \text{ in } x_1^n. \tag{1.3}$$

To summarize, the (conditional) distribution of the Markov chain $\{X_n\}$ is described by a proper *context set* T, and by a collection $\theta = \{\theta_s; s \in T\}$ of probability distributions $\theta_s = (\theta_s(0), \theta_s(1), \dots, \theta_s(m-1))$ for each element of the context set T.

The distribution of $\{X_n\}$ is determined as in (1.2), once we have specified a (proper) context set T – the *model* – and a collection $\theta = \{\theta_s; s \in T\}$ of probability vectors θ_s, for each $s \in T$ – the *parameters*. Note that the context set T can be described as a tree, a representation we will find very useful in the sequel. Therefore, we will refer to models T as context trees, context sets, or simply as models, interchangeably. In the tree representation, the context corresponding to the empty string λ is the root of the tree.

Example. Consider a 5th order Markov chain on the alphabet $A = \{0, 1, 2\}$, defined by the context tree T shown below, and by a collection of (known) parameters $\theta = \{\theta_s; s \in T\}$, where θ_s is a probability vector corresponding to leaf s in T.



Then the likelihood of an arbitrary string is easily computable explicitly via (1.1) or (1.2). For example, with $d = 5$ and $n = 12$, the string,

$$\underbrace{1,0,2,1,1,}_{x_{-4}^0}\underbrace{0,0,1,2,1,0,2,0,0,2,0,1}_{x_1^{12}}$$

has probability given by (1.1),

$$\theta_1(0)\cdot\theta_{01}(0)\cdot\theta_{00}(1)\cdot\theta_1(2)\cdot\theta_2(1)\cdot\theta_1(0)\cdot\theta_{01}(2)\cdot\theta_2(0)\cdot\theta_{0201}(0)\cdot\theta_{00}(2)\cdot\theta_2(0)\cdot\theta_{02002}(1).$$

Alternatively, we have the count vectors,

$$\begin{aligned}&a_1 = (2,0,1) \;\; a_2 = (2,1,0)\\ &a_{01} = (1,0,1) \;\; a_{00} = (0,1,1)\\ &a_{0201} = (1,0,0) \;\; a_{02002} = (0,1,0),\end{aligned}$$

with all other s having all-zero count vectors a_s, so that the probability of x_1^n given x_{-d+1}^0 as expressed in (1.2) is:

$$\theta_1(0)^2 \cdot \theta_1(2) \cdot \theta_2(0)^2 \cdot \theta_2(1) \cdot \theta_{01}(0) \cdot \theta_{01}(2) \cdot \theta_{00}(1) \cdot \theta_{00}(2) \cdot \theta_{0201}(0) \cdot \theta_{02002}(1).$$

Observe the convention that contexts s are written in the "backwards in time" direction, so that, for example, the fact that $a_{01}(2) = 1$ means that there is exactly one place in the string where a "2" follows the pattern "10."

Taking a Bayesian point of view, the natural next step is to define an appropriate prior structure on the class of models and the associated parameters. Given a fixed depth D and an arbitrary $\beta \in (0,1)$, we define a prior distribution on models (proper context sets, or the corresponding trees) T as,

$$\pi(T) = \pi_D(T) = \pi_D(T;\beta) = \alpha^{|T|-1}\beta^{|T|-L_D(T)}, \tag{1.4}$$

where $\alpha = (1-\beta)^{1/(m-1)}$, $|T|$ denotes the number of leaves of T, and $L_D(T)$ denotes the number of leaves T has at depth D.

Given a model (i.e., a context tree) T, we define a prior distribution on the probability vectors $\theta = \{\theta_s; s \in T\}$ on the leaves s of the context tree T: We place an independent Dirichlet$(1/2, 1/2, \ldots, 1/2)$ distribution on each θ_s so that, $\pi(\theta|T) = \prod_{s\in T} \pi(\theta_s)$, where,

$$\pi(\theta_s) = \pi(\theta_s(0), \theta_s(1), \ldots, \theta_s(m-1)) = \frac{\Gamma(m/2)}{\pi^{m/2}} \prod_{j=0}^{m-1} \theta_s(j)^{-\frac{1}{2}} \propto \prod_{j=0}^{m-1} \theta_s(j)^{-\frac{1}{2}}. \tag{1.5}$$

Finally, given the model T and the associated parameters $\theta = \{\theta_s; s \in T\}$, the likelihood of the observations is given as in (1.1) and (1.2),

$$P(x_1^n|x_{-d+1}^0) = P(x_1^n|x_{-d+1}^0, \theta, T) = \prod_{s\in T}\prod_{j=0}^{m-1} \theta_s(j)^{a_s(j)}, \tag{1.6}$$

where, as before, each $a_s(x)$ is the number of times x follows the context s in x_1^n. By convention, when we write $\sum_{s\in T}$ or $\prod_{s\in T}$, we take the corresponding sum or product over all the *leaves* s of the tree, not all its nodes. Also, in order to avoid cumbersome notation, in most of what follows we write x for the string x_1^n and we suppress the dependence on its initial context x_{-d+1}^0, so that, for example, we denote,

$$P(x, \theta|T) = P(x_1^n, \theta|x_{-d+1}^0, T).$$

An important and very useful property of this prior specification is that the parameters θ can easily be integrated out, so that the marginal likelihoods $P(T|x)$ can be expressed in closed form: The *marginal likelihood* $P(x|T)$ of the observations x given a model T is,

$$P(x|T) = \int P(x,\theta|T)d\theta = \int P(x|\theta, T)\pi(\theta|T)d\theta = \prod_{s\in T} P_e(a_s),$$

where the count vectors a_s are defined in (1.3) as before, and where the quantity $P_e(a)$ is given by,

$$P_e(a) = \frac{\prod_{j=0}^{m-1}[(1/2)(3/2)\cdots(a(j)-1/2)]}{(m/2)(m/2+1)\cdots(m/2+M-1)}, \tag{1.7}$$

for a count vector $a = (a(0), a(1), \ldots, a(m-1))$, where $M = a(0) + a(1) + \cdots + a(m-1)$, and with the convention that any empty product is taken to be equal to 1.

In terms of inference, the more interesting quantity is the model posterior distribution,

$$\pi(T|x) = \frac{P(x|T)\pi(T)}{P(x)}.$$

As usual, the main obstacle in the computation of $\pi(T|x)$ is the appearance of $P(x)$, which can be expressed as the weighted mean of the marginal likelihoods $P(x|T)$: $P(x) = \sum_T \pi(T)P(x|T)$. Since $P(x)$ will be central in much of the subsequent development, we refer to it as the *mean marginal likelihood* of the observations x.

2 Modeling and Sparse representation of nonlinear wireless channels

In this chapter we will review the basic models for the wireless channel nonlinearities. After providing a brief overview of the *linear* fading channel characteristics, a survey of past work on modeling of nonlinearities is presented. Subsequently, models for the description of nonlinear multi input multi output (MIMO) sources and systems are reviewed. Sparsity is a key constraint imposed on the model. The presence of sparsity is often dictated by physical considerations as in wireless fading channel–estimation. In other cases it appears as a pragmatic modelling approach that seeks to cope with the curse of dimensionality, particularly acute in nonlinear systems like Volterra type series. When system nonlinearities are present, possible remedies based on linear approximations may degrade system performance. A popular model that captures system nonlinearities is the Volterra series [73, 74, 75]. This model is employed in many applications including wireless communications. Volterra series constitute a class of polynomial models that can be regarded as a Taylor series with memory. An attractive feature of this model is that the unknown parameters enter linearly at the output. On the other hand, the number of terms increases exponentially with the order and memory of the model. Most of the work reported in the literature focuses on modelling and identification of single input single output (SISO) Volterra systems. When the underlying nonlinear system is a MIMO system, as in MIMO communications, the resulting model is more complicated and has received little attention. MIMO models are addressed in this chapter. Nonlinear MIMO systems involve a large number of parameters, which increases exponentially with the order, the memory and the number of inputs. Therefore, there is a strong need to reduce complexity by considering those terms that strongly contribute to the outputs. This leads naturally to a sparse approximation of the underlying nonlinear MIMO system. The models described in this chapter will be used in the estimation of sparse nonlinear MIMO sources and channels.

2.1 Modeling of wireless fading channels

The wireless signal transmission between two points for telecommunication links can be realized either using radio frequencies (RF), or optical frequencies. In the latter case the systems are known as optical wireless or free space optical (FSO) communications systems.

2.1.1 The RF Wireless Channel

The RF wireless channel models RF propagation for the purposes of mobile antennas in the presence of significant scattering. Typically, the models incorporate three generic types of cellular environments: suburban macrocells, urban macrocells, and urban microcells. The relationship between a given channel scenario and the channel coefficients for a given link can be described in terms of three levels of abstraction [53].

At the macroscopic level, time-averaged local properties of the channel are described, e.g., the average power, delay spread (DS), and angle spread (AS). These quantities are also designated as "composite" parameters to imply the inclusion of all delayed components. Apart from a deterministic part, these variables have a log-normal random part, which captures the fluctuations due to propagation through several independent "city block" regions.

Focusing in to a deeper "mesoscopic" level, the channel has additional structure. In particular, each composite energy cluster is decomposed into multiple paths with relative delays, and angles of arrival and angles of departure consistent with the composite statistics. Each of these paths can be thought of as coming from different buildings within the neighborhood of that block. Also at this mesoscopic level, the path delays and average path powers are generated as realizations of random variables.

At the deepest, microscopic level, each of these paths undergoes Rayleigh fading, generated from the temporal variability of the particular link (e.g., due to the terminal's movement). Each path is represented as a sum of subpaths modeled as planewaves. Since the various length-scales are not always clearly separable, the interpretation of these levels of abstraction does not always correspond with reality. However, they certainly make sense and can always be used to describe the experimental data of outdoors channels. In any event, the above characteristics are intertwined with the need of power amplifiers, which inevitably produce nonlinearities. These have to be studied together and not separately from the above (linear) characteristics of the channel.

2.1.2 The FSO Channel

On the other hand, the FSO links use optical wavelengths and more specifically operate in the range between $0.68\mu m$—$1.55\mu m.$ In these wavelengths the atmospheric attenuation is strong and thus the maximum operational range of these systems is about five or six km, depending on the atmospheric conditions between the transmitter and the receiver. Moreover, this communication link demands full line of sight between the transmitter and the receiver of the point to point link. However, the FSO links can achieve very high bandwidth, secure communications, with low interference and relatively low installation and operational cost. Additionally, this technology does not need any license to establish a new link. Furthermore, the performance of the FSO links depends strongly on the atmospheric conditions in the area of link. Thus, fog, strong rain and hail can mitigate its performance. Moreover, the atmospheric turbulence conditions are decreasing the channel's performance due to the scintillation effect which transforms the optical static channel into a fading one.

The fading may vary in time, or in frequency and is modeling as a random process and can be caused either due to the multiple propagation paths, either due to shadowing from obstacles either from turbulence effect, mostly in very small wavelengths referring to optical links. Fading channels often, are modeling the effects of electromagnetic wave transmission in the atmosphere and usually are modeling as time-varying random process which affects the amplitude and the phase of the information signal. The fading channel could follow either fast of slow fading statistics. The slow fading resulting when the coherence time of the channel is relative to its delay constraint , [54, 55]. In this case the changes, caused by the fading propagation path, in amplitude and phase of the information signal can, accurately considered as constant for a period of use.

Many statistical models have been proposed in order to model accurately the fading effect. It is obvious that their accuracy depends on the operational wavelength of the emulated wireless communication system, as well, the obstacles intruding inside the propagation path, the multipath selection, the atmospheric conditions, the strength of the turbulence effect, etc. Thus, Rayleigh and the Rice are very often using model, while the log normal, the Weibull, the Nakagami, the K, the I-K, the negative exponential and the gamma gamma distribution are providing accurately distribution model for specific, realistic enough, setups [56, 57, 58, 59, 60, 61, 62]. As a result, optical amplifiers are necessary for overcoming this decreasing. Therefore, their existence in the FSO link, adds nonlinearities at the whole system's performance evaluation.

2.2 Models for wireless channel nonlinearities

2.2.1 Volterra Models

Volterra series constitute a popular model for the description of nonlinear behaviour [73, 74]. A single-input single-output (SISO) discrete–time Volterra model has the following form

$$y(n) = \sum_{p=1}^{\infty} \sum_{\tau_1=-\infty}^{\infty} \cdots \sum_{\tau_p=-\infty}^{\infty} h_p(\tau_1, \ldots, \tau_p) \left[\prod_{i=1}^{p} x(n-\tau_i) \right] . \tag{2.1}$$

Each output is formed by weighting the input shifted samples $x(n-\tau_i)$ and their products. The weights $h_p(\tau_1,\ldots,\tau_p)$ constitute the *Volterra kernels* of order p. Well posed conditions ensuring that inputs give rise to well defined outputs are given in [135, 136]. If only a finitely number of nonlinearities enters Eq. (2.1), the resulting expression defines a finite Volterra system. Suppose the kernels of a finite Volterra system are causal and absolutely summable. Then Eq. (2.1) defines a bounded input bounded output (BIBO) stable system and can be approximated by the polynomial system

$$y(n) = \sum_{p=1}^{P}\sum_{\tau_1=0}^{M}\cdots\sum_{\tau_p=0}^{M} h_p(\tau_1,\ldots,\tau_p)\left[\prod_{i=1}^{p} x(n-\tau_i)\right]. \tag{2.2}$$

Eq. (2.2) is parametrized by the finite Volterra kernels and has finite memory M. A more general result established by Boyd and Chua [137, 136] states that any shift invariant causal BIBO stable system with *fading memory* can be approximated by Eq. (2.2). The fading memory is a continuity property with respect to a weighted norm which penalizes the remote past in the formation of the current output. The reader may consult [137, 136, 135] for more details.

A key feature of Eq. (2.2) is that it is *linear in the parameters*. For estimation purposes it is useful to write Eq. (2.2) in matrix form using Kronecker products [138]. Indeed, let $\vec{x}(n) = [x(n), x(n-1),\cdots,x(n-M)]^T$ and the pth-order Kronecker power

$$\vec{x}_p(n) = \underbrace{\vec{x}\otimes\cdots\otimes\vec{x}}_{p \text{ times}}, \quad p = 2,\ldots,P.$$

The Kronecker power contains all pth–order products of the input. Likewise $\vec{h} = \left[\vec{h}_1(\cdot),\cdots,\vec{h}_p(\cdot)\right]^T$ is obtained by treating the p–dimensional kernel as a M^p column vector. We rewrite Eq. (2.2) as follows

$$y(n) = \left[\vec{x}^T(n)\cdots\vec{x}_p^T(n)\right]\begin{bmatrix}\vec{h}_1\\ \vdots\\ \vec{h}_p\end{bmatrix} = \vec{x}^T(n)\vec{h}. \tag{2.3}$$

Collecting n successive output samples from the above equation into the vector $\vec{y}(n) = [y(1),\ldots,y(n)]$ results in the following system of linear equations:

$$\vec{y}(n) = \vec{X}(n)\vec{h}$$

when

$$\vec{X}(n) = \left[\vec{x}^T(1),\ldots,\vec{x}^T(n)\right]^T.$$

From a practical viewpoint, Volterra models of order higher than three are rarely considered. This is due to the fact that the number of parameters involved in the model of Eq. (2.2) grows exponentially as a function of the memory size and the order of nonlinearity (#parameters: $\sum_{p=1}^{P} M^p$). To cope with this complexity several sub–families of Eq. (2.2) have been considered, most notable Wiener, Hammerstein and Wiener–Hammerstein models. In all cases the universal approximation capability is lost.

A Wiener system is the cascade of a linear filter followed by a static nonlinearity. If we approximate the static nonlinearity with its Taylor expansion up to a certain order, we obtain the following expression for the output of the Wiener system

$$y(n) = \sum_{p=1}^{P}\left[\sum_{\tau=0}^{M} h_p(\tau)x(n-\tau)\right]^p. \tag{2.4}$$

The Hammerstein system (or memory polynomial) is composed of a memoryless nonlinearity (a Taylor approximation of the static nonlinearity is employed) followed by a linear filter, and has the following form

$$y(n) = \sum_{p=1}^{P}\sum_{\tau=0}^{M} h_p(\tau)x^p(n-\tau). \tag{2.5}$$

A Wiener–Hammerstein or sandwich model is composed of a memoryless nonlinearity sandwiched between two linear filters with impulse responses $h(\cdot)$ and $g(\cdot)$ and is defined as

$$y(n) = \sum_{p=1}^{P} \sum_{\tau_1=0}^{M} \cdots \sum_{\tau_p=0}^{M} \sum_{k=0}^{M_{h_p}+M_{g_p}} g_p(k) \prod_{l=1}^{p} h_p(\tau_l - k) x(n - \tau_l). \tag{2.6}$$

The above models have been employed in a wide range of applications including: satellite, telephone channels, mobile cellular communications, wireless LAN devices, radio and TV stations, digital magnetic systems and others [139, 73, 75, 140, 141].

2.2.2 *Models for power clipping*

The above Volterra-type approach provides a good description on the low-nonlinearity region of power amplification. To describe the region where the amplification leads to clipping, i.e. to the cutoff of higher power outputs, one needs to be able to model the larger output power regions. One way to do this it to provide models for modulation of the output phase due to amplitude non-linearities, represented by amplitude modulation (AM) to phase modulation (PM) conversion (AM/PM) and the amplitude modulation to amplitude modulation conversion (AM/AM) due to the nonlinearity. The first conversion refers to the amount of undesired phase deviation that is caused by amplitude variations of the system while the second one to undesired amplitude deviation [68]. A simple way to represent this behavior is given by Saleh's model. Specifically, the AM/AM distortion is given by the following equation [70]:

$$A(r) = \frac{a_A r}{1 + b_A |r|^2} \tag{2.7}$$

and the AM/PM distortion by the equation

$$P(r) = \frac{a_P r}{1 + b_P |r|^2} \tag{2.8}$$

where a and b are the parameters that characterize the behavior of the model and r(t) is the amplitude of the input signal.

Other models that deal with strong nonlinearities are the following [72]:

- Soft limiter. This model can approach the physical behavior of an amplifier in case of using a suitable predistorter which will linearize the nonlinear element.

$$F[x] = \min(|x|, A) \tag{2.9}$$

- Solid State Power amplifier.The relationship between input and output is given by

$$F_p[x] = \frac{x}{\left[1 + \left(\frac{|x|}{A}\right)^{2p}\right]^{\frac{1}{2p}}} \tag{2.10}$$

In this model there is a parameter,p , which controls the smoothness of the transition from the linear region to the saturation region.

2.3 Linearly Mixed models

In MIMO systems the signals from the n_i inputs interact with each other and the resulting mixture is received at each output. If the path between each input and each output is modelled as a Volterra system, then the rth output is expressed as follows

$$y_r(n) = \sum_{p=1}^{P} \sum_{t=1}^{n_i} \sum_{\tau_1=0}^{M} \cdots \sum_{\tau_p=0}^{M} h_p^{(r,t)}(\tau_1, \ldots, \tau_p) \prod_{i=1}^{p} x_t(n - \tau_i) \tag{2.11}$$

where $h_p^{(r,t)}(\tau_1, \ldots, \tau_p)$ is the *pth–order Volterra kernel* between the tth input and the rth output for all $t = 1, \ldots, n_i$ and $r = 1, \ldots, n_o$. The above model does not allow product combinations along different inputs. Instead each input is nonlinearly transformed and then all different inputs are linearly mixed. Such a model can be considered as a parallel cascade of n_i SIMO Volterra models.

We start by defining the tth input regressor vector as

$$\vec{x}^{(t)}(n) = [x^{(t)}(n), x^{(t)}(n-1), \ldots, x^{(t)}(n-M)]^T$$

and thus the linearly mixed input vector is defined in the following compact way:

$$\vec{x}(n) = [\vec{x}_1^{(1)}(n), \vec{x}_2^{(1)}(n), \ldots, \vec{x}_p^{(1)}(n), \cdots, \vec{x}_1^{(n_i)}(n), \vec{x}_2^{(n_i)}(n), \ldots, \vec{x}_p^{(n_i)}(n)]^T.$$

The total number of parameters of the above parallel cascade or linearly mixed model is

$$\#\text{parameters: } n_i \sum_{i=1}^{p} M^p$$

and is considerably reduced when compared to the general case.

The linearly mixed model finds application in nonlinear communications. Communication nonlinearities can be categorized into the following three types: transmitter nonlinearity (due to nonlinearity in amplifiers), inherent physical channel nonlinearity, and receiver nonlinearity (e.g., due to nonlinear filtering). The Power Amplifier (PA) (which is located at the transmitter) constitutes the main source of nonlinearity for several communication systems. In a system equipped with multiple transmit antennas, each transmitter amplifies the signal. Amplifiers often operate near saturation to achieve power efficiency. In those cases they introduce nolinearities which cause interference and reduce spectral efficiency. At the receiver end, each antenna receives a linear superposition of all transmitted signals, as illustrated in Fig. 2.1. It should be pointed out that the nonlinear effects are applied to each input signal individually prior to mixing the transmitted signals. Finally, it should also be stressed that, since amplifiers typically jointly amplify signals for different users, the interference caused through the non-linearities should also be taken into account.

2.4 Nonlinear communication systems

MIMO communication systems equipped with multiple transmit and/or receive antennas are MIMO systems that help provide spatial diversity. Exploitation of spatial diversity results in higher capacity and performance improvements in interference reduction, fading mitigation and spectral efficiency. Most of existing MIMO schemes are limited to linear systems. However, in many cases, system nonlinearities are present and possible remedies based on linear MIMO approximations degrade performance significantly.

In a communication system, there are often limited resources (power, frequency, and time slots) which have to be efficiently shared by many users. Quite often in practice we encounter a situation whereby the number of users exceeds the number of available frequency or time slots. In infrastructure–based networks, a base station or an access point is responsible for simultaneously sharing the resources among the users, thereby reducing the access delays/transmission latency and improving quality–of–service (QoS). This is established through a variety of *multiple access* schemes. Two key multiple access technologies suitable for higher data rates and performance are: orthogonal frequency–division multiple access (OFDMA) and code–division multiple access (CDMA).

OFDMA is a popular multiple access method, for high–speed communications, whereby it dynamically allocates resources both in frequency (by dividing the available bandwidth into a number of subbands, called subcarriers) and in time (via OFDM symbols). The transmission assigns different users to groups of orthogonal subcarriers and thus allows them to be spaced very close together with no overhead as in frequency division multiple access. Furthermore it prevents interference between adjacent subcarriers. OFDMA has been implemented in several wireless communication standards (IEEE 802.11a/g/n

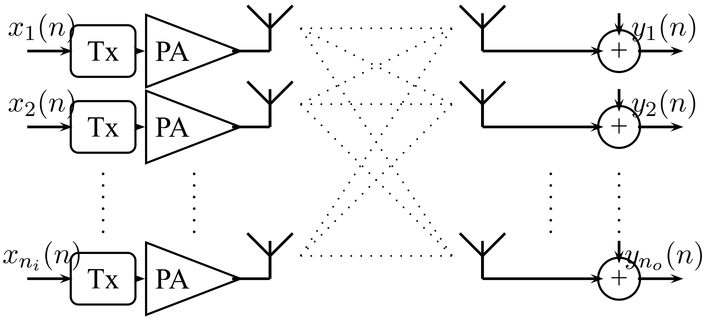


Figure 2.1: An example of a parallel cascade MIMO Volterra channel

wireless local area networks (WLANs), IEEE 802.16e/m worldwide interoperability for microwave access (WiMAX), Hiperlan II), high–bit–rate digital subscriber lines (HDSL), asymmetric digital subscriber lines (ADSL), very high-speed digital subscriber lines (VHDSL), digital audio broadcasting (DAB), digital television and high-definition television (HDTV).

OFDMA is capable of mitigating intersymbol interference (ISI), due to multipath propagation) using low–complexity/simple equalization structures. This is established by transforming the available bandwidth into multiple orthogonal narrowband subcarriers, where each subcarrier is sufficiently narrow to experience relatively flat fading. Nevertheless, OFDM is sensitive to synchronization issues and is characterized by high peak–to–average–power–ratio (PAPR), caused by the sum of several symbols with large power fluctuations. Such variations are problematic because practical communication systems are peak powered limited. In addition, OFDM transceivers are also intrinsically sensitive to power amplifier (PA) nonlinear distortion [143], which dissipates the highest amount of power. One way to avoid nonlinear distortion is to operate the PA at the so–called “back–off” regime which results in low power efficiency. The trade–off between power efficiency and linearity motivated the development of signal processing tools that cope with MIMO–OFDM nonlinear distortion [144, 145, 143].

CDMA is based upon spread spectrum techniques which have been used by the military for decades. Spread spectrum techniques play an important role in third generation mobile systems (3G) and have found application in IEEE 802.11b/g (WLAN), Bluetooth, and cordless telephony. In CDMA multiple users share the same bandwidth at the same time through the use of (nearly) orthogonal spreading codes. The whole process effectively spreads the bandwidth over a wide frequency range (using pseudo-random code spreading or frequency hopping) several magnitudes higher than the original data rate.

Two critical factors that limit the performance of CDMA systems are interchip/intersymbol interference (ICI/ISI), due to multipath propagation, mainly because they tend to destroy orthogonality between user codes and thus prevent interference elimination. Suppression of the detrimental effects of interference (ICI and ISI) get further complicated when nonlinear distortion is introduced due to power amplifiers. The combined effects of ICI, ISI and nonlinearities are comprehensively examined in [146, 144]. However, as recently illustrated in [147] the CDMA system model is sparse due to user inactivity/uncertainty, timing offsets and multipath propagation. CDMA system performance can be expected to improve further if nonlinearities along with sparse ICI/ISI are revisited.

3 Research Issues for SWINCOM and conclusions

The following research issues will be addressed in the context of representation and estimation of non–linear sources.

a. **Development and estimation of sparse locally affine manifold models.** As mentioned earlier, CS involves linear and data-nonadaptive operators for compression and reconstruction. A recent noteworthy effort to depart from the linear CS paradigm and develop linear-quadratic CS operators can be found under the term quadratic basis pursuit in [34]. However, this approach too is data-nonadaptive, which means that the (de-) coding operators are "one-size-fits-all" random matrices that satisfy generalized restricted isometry properties, but totally ignore the underlying signal statistics. As a result, they cannot even come close to jpeg and mpeg standard (de-) coding modules.

   Our first approach toward nonlinear and data-adaptive (de-) coders will rely on the sparse (locally) affine manifold models we alluded to in the previous subsection [40]. The framework will pursue a two-pronged objective: (a) Based on $\{\boldsymbol{x}_n\}_{n=1}^N$, the goal is to develop a sparsity-aware, outlier-resilient estimate of the manifold $\mathcal{M}$, and map it to a lower-dimensional space $\mathbb{R}^d$ (with $d \ll D$) by *robust sparse embeddings* in the training phase; and (b) leverage this mapping during the operational phase to "compress" $\boldsymbol{x} \in \mathbb{R}^D$ as $\boldsymbol{y} \in \mathbb{R}^d$ at the Tx, and use the latter (or its noisy version $\hat{\boldsymbol{y}}$) to reconstruct an estimate $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ at the Rx [40].

   We will derive novel performance bounds to assess performance of these (locally) affine regression-type and bilinear DL-type (de-)coders, both analytically and with thorough testing of simulated and real image and audio data. We will further gauge performance of our novel schemes in clustering, classification, sampling, interpolation, extrapolation, and reconstruction even when the source waveforms have missing samples.

b. **Development and estimation of non–linear PCA filterbanks and joint design of PCA filterbanks and quantizers.** The second approach we will pursue is through nonlinear PCA filterbanks. Those will build on our prior works on linear PCA-based filterbanks in [43]. The potential of this approach is corroborated by the success we had in equalizing Volterra communication channels in [21].

   We will jointly design these nonlinear and data-adaptive dimensionality reduction steps with (vector) quantizers [23]. In addition to investigating novel quantizer designs, it will be important to explore optimal reconstruction schemes from quantized vectors. Recent efforts to reconstructing CS vectors using quantized data can be found in e.g., [7, 8, 26, 29, 30]. Reconstruction performance from quantized measurements for DL and for the novel (de-)coders proposed here is an uncharted territory, and thus a fertile ground for exciting research.

c. **Estimation and compression of sparse discrete sources.** New techniques will be developed for the estimation and compression of sparse discrete sources. Rissanen tree sources, variable memory Markov Chains and context tree maximization procedures will form the foundation of our proposed development.

d. **Approximations of non–linear MIMO systems with Universal approximation capability.** The aim is to develop finitely parametrizable structures that extend the single input models to multi input multi output settings and poses a universal approximation capability. Furthermore, parsimonious models will be identified, using sparsity considerations. These will include MIMO Wiener–Hammerstein models.

List of figures

List of tables

List of algorithms

Bibliography

1. M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
2. A. D’Aspremont, F. Bach, and L. E. Ghaoui, “Optimal solutions for sparse principal component analysis,” *Journal of Machine Learning Research*, vol. 9, pp. 1269–1294, 2008.
3. R. G. Baraniuk, E. Candes, R. Nowak, and M. Vetterli, “Compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 12–13, 2008.
4. R. G. Baraniuk and M. B. Wakin, “Random projections of smooth manifolds,” *Found. of Comput. Math.*, vol. 9, no. 1, pp. 51–77, Feb. 2009.
5. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Prentice Hall, 1971.
6. D. P. Bertsekas, *Nonlinear Programming*. Second Edition, Athena Scientific, 2003.
7. P. T. Boufounos and R. G. Baraniuk, “1-bit Compressive Sensing,” in *Proc. Conf. on Information Science and Systems*, Princeton, NJ, Mar. 2008.
8. P. T. Boufounos, “Greedy sparse signal reconstruction from sign measurements, in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, Asilomar, CA, Nov. 2009.
9. D. R. Brillinger, *Time Series: Data Analysis and Theory*. Expanded Edition, Holden Day, 1981.
10. E. Candès and T. Tao, “Near optimal signal recovery from random projections: Universal encoding strategies?” *IEEE Trans. on Info. Theory*, pp. 5406–5425, Dec. 2006.
11. E. Candes and B. Recht, “Exact matrix completion via convex optimization,” *Found. Comp. Math.*, vol. 9, pp. 717-772, 2008.
12. G. Cao and C. A. Bouman, “Covariance estimation for high-dimensional data vectors using the sparse matrix transform,” *Proc. of Neural Info. Proc. Systems Conf.*, December 2008.
13. L. Carin, R. G. Baraniuk, V. Cevher, D. Dunson, M. I. Jordan, G. Sapiro, and M. B. Wakin, “Learning low-dimensional signal models,” *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 39–51, Mar. 2011.
14. J. Chen, X. Zhang, T. Berger and S. B. Wicker, “An Upper Bound on the Sum-Rate Distortion Function and Its Corresponding Rate Allocation Schemes for the CEO Problem,” *IEEE Journal on Sel. Areas in Comm.*, vol. 50, pp. 406–411, August 2004.
15. M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin, “Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds,” *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6140–6155, Dec. 2010.
16. S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Review*, vol. 43, no. 1, pp. 129–159, 1998.
17. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Ed., John Wiley and Sons, 1991.
18. D. L. Donoho, “Compressed Sensing,” *IEEE Trans. on Info. Theory*, vol. 52, pp. 1289-1306, 2006.
19. E. Elhamifar and R. Vidal, “Sparse manifold clustering and embedding,” in *Proc. of Neural Inf. Process. Syst.*, Granada: Spain, Dec. 2011.
20. P. Forero and G. B. Giannakis, “Sparsity-exploiting robust multidimensional scaling,” *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4118–4134, Aug. 2012.
21. G. B. Giannakis and E. Serpedin, “Linear multichannel blind equalizers of nonlinear FIR Volterra channels,” *IEEE Trans. on Signal Processing*, vol. 45, pp. 67-81, January 1997.
22. A. Ghodsi, “Dimensionality reduction: A short tutorial,” University of Waterloo, Tech. Rep. 2006-14, 2006.
23. R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2325-2383, Oct. 1998.
24. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.
25. S. Huang, C. Cai, and Y. Zhang, “Dimensionality reduction by using sparse reconstruction embedding,” in *Lecture Notes in Computer Science*, vol. 6298, 2010, pp. 167–178.

26. L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, “Robust 1-Bit Compressive Sensing via Binary Stable Embeddings of Sparse Vectors,” Apr. 2011. Available: url=http://arxiv.org/abs/1104.3160

27. I. Jolliffe, *Principal Component Analysis*, Second Edition, New York: Springer, 2002.

28. D. Kong, C. Ding, H. Huang, and F. Nie, “An iterative locally linear embedding algorithm,” in *Proc. of the Int. Conf. on Machine Learning*, 2012, url=http://arxiv.org/abs/1206.6463.

29. J. N. Laska, P. Boufounos, M. A. Davenport, and R. G. Baraniuk, “Democracy in Action: Quantization, Saturation, and Compressive Sensing,” *Applied and Computational Harmonic Analysis*, v. 31, no. 3, pp. 429-443, Nov. 2011.

30. J. N. Laska, Z. Wen, W. Yin, and R. G. Baraniuk, “Trust, but Verify: Fast and Accurate Signal Recovery from 1-bit Compressive Measurements,” *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5289–5301, Nov. 2011.

31. Z. Lu and Y. Zhang, “An augmented Lagrangian approach for sparse principal component analysis,” 2009. [Online]. Available: url=http://www.citebase.org/abstract?id=oai:arXiv.org:0907.2079

32. J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online dictionary learning for sparse coding,” in *Proc. of the 26th Intl. Conf. on Machine Learning*, Montreal, Canada, 2009, pp. 689–696.

33. G. Mateos and G. B. Giannakis, “Robust PCA as Bilinear Decomposition with Outlier-Sparsity Regularization,” *IEEE Trans. on Signal Processing*, vol. 60, no. 10, pp. 5176-5190, October 2012.

34. H. Ohlsson, A. Y. Yang, R. Dong, M. Verhaegen, S. Sastry, “Quadratic Basis Pursuit,” arXiv:1301.7002 [cs.IT], 2013.

35. B. A. Olshausen and D. J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by V1?” *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

36. D. J. Sakrison, “Source encoding in the presence of random disturbance,” *IEEE Trans. on Info. Theory*, vol. 14, pp. 165–167, January 1968.

37. L. K. Saul and S. T. Roweis, “Think globally, fit locally: Unsupervised learning of low dimensional manifolds,” *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

38. I. D. Schizas and G. B. Giannakis, “Covariance Eigenvector Sparsity for Compression and Denoising,” *IEEE Trans. on Signal Processing*, vol. 60, no. 5, pp. 2408-2421, May 2012.

39. B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2001.

40. K. Slavakis, G. B. Giannakis, and G. Leus, “Robust Sparse Embedding and Reconstruction via Dictionary Learning,” *Proc. of Conf. on Info. Sciences and Systems*, John Hopkins Univ., March 2013.

41. R. Timofte and L. van Gool, “Sparse representation based projections,” in *Proc. of the British Machine Vision Conf.*, 2011.

42. I. Tošić and P. Frossard, “Dictionary learning,” *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 27–38, Mar. 2011.

43. M. K. Tsatsanis and G. B. Giannakis, “Principal component filter banks for optimal multiresolution analysis,” *IEEE Trans. on Signal Processing*, vol. 43, no. 8, pp. 1766-1777, August 1995.

44. M. O. Ulfarsson and V. Solo, “Sparse variable PCA using geodesic steepest descent,” *IEEE Trans. on Signal Processing*, vol. 10, no. 12, pp. 5823–5832, 2008.

45. ——, “Sparse variable PCA using a steepest descent on a Grassman manifold,” in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 3253–3256.

46. H. S. Witsenhausen, “Indirect Rate Distortion Problems,” *IEEE Trans. on Info. Theory*, vol. 26, pp. 518–521, September 1980.

47. J. Wolf and J. Ziv, “Transmission of noisy information to a noisy receiver with minimum distortion,” *IEEE Trans. on Info. Theory*, vol. 16, pp. 406–411, July 1970.

48. J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. and Machine Intelligence*, vol. 31, no. 2, pp. 1–18, Feb. 2009.

49. A. Wyner and J. Ziv, “The Rate-Distortion Function for Source Coding with Side Information at the Decoder,” *IEEE Trans. on Info. Theory*, vol. 26, pp. 1–10, January 1976.

50. H. L. Yap, M. B. Wakin, and C. J. Rozell, “Stable manifold embeddings with operators satisfying the restricted isometry property,” in *Proc. of 45th Conf. on Info. Sciences and Systems*, Baltimore: Maryland, Mar. 2011.

51. ——, “Stable manifold embeddings with structured random matrices,” url=http://arxiv.org/abs/1209.3312, 2012.

52. H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, 2006.

53. G. Calcev, D. Chizhik, B. Goeransson, S. Howard, H. Huang, A. Kogiantis, A. F. Molisch, A. L. Moustakas, D. Reed, and H. Xu, “A wideband spatial channel model for system-wide simulations,” *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, p. 389, Mar. 2007.

54. T. S. Rappaport, *Wireless Communications: Principles and practice*, 2nd ed. Prentice Hall, 2002.

55. D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

56. L. C. Andrews, *Atmospheric optics*. SPIE Optical Engineering Press, 2004.

57. H. E. Nistazakis, E. A. Karagianni, A. D. Tsigopoulos, M. E. Fafalios, and G. S. Tombras, “Average capacity of optical wireless communication systems over atmospheric turbulence channels,” *Journal of Lightwave Technology*, vol. 27, pp. 974–979, 2009.

58. L. C. Andrews and R. L. Philips, “I-k distribution as a universal propagation model of laser beams in atmospheric turbulence,” *Journal of the Optical Society of America*, vol. 2, pp. 160–163, 1985.

59. H. E. Nistazakis, V. D. Assimakopoulos, and G. S. Tombras, “Performance estimation of free space optical links over negative exponential atmospheric turbulence channels,” *Journal of light and electron optics*, January 2011.

60. H. Henniger and O. Wilfert, “An introduction to free-space optical communications,” *Radioengineering*, vol. 19, pp. 203–212, 2010.

61. A. K. Majumdar, “Free-space laser communication performance in the atmospheric channel,” *Journal of Optical Fiber Communications*, pp. 345–396, 2005.

62. A. Stassinakis, H. Nistazakis, and G. Tombras, “Comparative performance study of one or multiple receivers schemes for fso links over gamma–gamma turbulence channels,” *Journal of Modern Optics*, vol. 59, pp. 1023–1031, 2012.

63. J. Gomez and E. Baeyens, “Hammerstein and wiener model identification using rational orthonormal.”

64. G. Budura, “Nonlinear systems identification using the volterra model.”

65. T. Liu, S. Boumaiza, and F. M. Ghanouchi, “Deembedding static nonlinearities and accurately identifying and modeling memory effects in wide-band RF transmitters,” *IEEE transactions on microwave theory and techniques*, vol. 53, pp. 3578–3587, November 2005.

66. D. Mirri, G. Iuculano, F. Filicori, G. Pasini, G. Vannini, and G. P. Gualtieri, “A modified volterra series approach for nonlinear dynamic systems modeling,” *IEEE Transactions on circuits and systems*, vol. 49, pp. 1118–1128, August 2002.

67. E. Eskinat, S. H. Johnson, and W. L. Luyben, “Use of hammerstein models in identification of nonlinear systems,” *AIChe Journal*, vol. 37, pp. 255–268, February 1991.

68. Karkahaneh, A. Ghorbani, and H. Amindavar, “Modeling and compensating memory effect in high power amplifier for ofdm systems,” *Progress In Electromagnetics Research C*, vol. 3, pp. 183–194, 2008.

69. A. Mittal and P. Aadaleesan, “A new hammerstein model for non-linear system identification,” *International Journal of Communication Network and Security (IJCNS)*, vol. 1.

70. E. Aschabacher and M. Rupp, “Modelling and identification of a nonlinear power-amplifier with memory for nonlinear digital adaptive pre-distortion,” in *IEEE Proceeding of the SPAWC Workshop*, june 2003.

71. A. Zhu and T. J. Brazil, “Nonlinear amplifier effects in communications systems,” *IEEE microwave and wireless components letters*, vol. 14, pp. 563–565, December 2004.

72. J. Tellado, L. Hoo, and J. Cioffi, “Maximum-likelihood detection of nonlinearly distorted multicarrier symbols by iterative decoding,” *IEEE Trans. Commun.*, vol. 51, no. 2, pp. 218–228, Feb. 2003.

73. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*. Willey and Sons, 1980.

74. N. S. Theory, *W.J. Rugh*. The Johns Hopkins University Press, 1981.

75. G. S. V.J. Mathews, *Polynomial Signal Processing*. Wiley-Blackwell, 2000.

76. J. Kieffer, “A survey of the theory of source coding with a fidelity criterion,” *IEEE Trans. Inform. Theory*, vol. 39, no. 5, pp. 1473–1490, 1993.

77. J. Ziv, “Coding of sources with unknown statistics – Part II: Distortion relative to a fidelity criterion,” *IEEE Trans. Inform. Theory*, vol. 18, no. 3, pp. 389–394, 1972.

78. R. Neuhoff, D.L. Gray and L. Davisson, “Fixed rate universal block source coding with a fidelity criterion,” *IEEE Trans. Inform. Theory*, vol. 21, no. 5, pp. 511–523, 1975.

79. J. Ziv, “Distortion-rate theory for individual sequences,” *IEEE Trans. Inform. Theory*, vol. 26, no. 2, pp. 137–143, 1980.

80. D. Ornstein and P. Shields, “Universal almost sure data compression,” *Ann. Probab.*, vol. 18, pp. 441–452, 1990.

81. J. Muramatsu and F. Kanaya, “Distortion-complexity and rate-distortion function,” *IEICE Trans. Fundamentals*, vol. E77-A, pp. 1224–1229, 1994.

82. Z. Zhang and V. Wei, “An on-line universal lossy data compression algorithm by continuous codebook refinement – Part I: Basic results,” *IEEE Trans. Inform. Theory*, vol. 42, no. 3, pp. 803–821, 1996.

83. Z. Zhang and E.-H. Yang, “An on-line universal lossy data compression algorithm by continuous codebook refinement – Part II: Optimality for phi-mixing models,” *IEEE Trans. Inform. Theory*, vol. 42, no. 3, pp. 822–836, 1996.

84. E.-H. Yang and J. Kieffer, “Simple universal lossy data data compression schemes derived from the Lempel-Ziv algorithm,” *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 239–245, 1996.

85. A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers, 1992.

86. T. Linder, G. Lugosi, and K. Zeger, “Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding,” *IEEE Trans. Inform. Theory*, vol. 40, no. 6, pp. 1728–1740, 1994.

87. P. Chou, M. Effros, and R. Gray, “A vector quantization approach to universal noiseless coding and quantizations,” *IEEE Trans. Inform. Theory*, vol. 42, no. 4, pp. 1109–1138, 1996.

88. R. Gray and D. Neuhoff, “Quantization,” *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

89. F. Jelinek, “Tree encoding of memoryless time-discrete sources with a fidelity criterion,” *IEEE Trans. Information Theory*, vol. IT-15, pp. 584–590, 1969.

90. J. Anderson and F. Jelinek, “A 2-cycle algorithm for source coding with a fidelity criterion,” *IEEE Trans. Information Theory*, vol. IT-19, no. 1, pp. 77–92, 1973.

91. A. Viterbi and J. Omura, “Trellis encoding of memoryless discrete-time sources with a fidelity criterion,” *IEEE Trans. Information Theory*, vol. IT-20, pp. 325–332, 1974.

92. R. Gray, “Time-invariant trellis encoding of ergodic discrete-time sources with a fidelity criterion,” *IEEE Trans. Information Theory*, vol. IT-23, no. 1, pp. 71–83, 1977.

93. M. Marcellin and T. Fischer, “Trellis coded quantization of memoryless and Gauss-Markov sources,” *IEEE Trans. Comm.*, vol. 38, no. 1, pp. 82–93, 1990.

94. R. van der Vleuten and J. Weber, “Construction and evaluation of trellis-coded quantizers for memoryless sources,” *IEEE Trans. Information Theory*, vol. 41, no. 3, pp. 853–859, 1995.

95. E. Martinian and M. Wainwright, “Low density codes achieve the rate-distortion bound,” in *Proc. Data Compression Conf. – DCC 2006*, Snowbird, UT, March 2006, pp. 153–162.

96. M. Wainwright and E. Maneva, “Lossy source encoding via message-passing and decimation over generalized codewords of LDGM codes,” in *Proc. of the IEEE International Symposium on Inform. Theory*, Adelaide, Australia, Sept. 2005, pp. 1493–1497.

97. I. Kontoyiannis, K. Rahnama Rad, and S. Gitzenis, “Superposition codes for Gaussian vector quantization,” in *IEEE Inform. Theory Workshop*, Cairo, Egypt, January 2010.

98. R. Venkataramanan, A. Joseph, and S. Tatikonda, “Gaussian rate-distortion via sparse linear regression over compact dictionaries,” in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, July 2012, pp. 368–372.

99. A. Barron and A. Joseph, “Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity,” *IEEE Trans. Inform. Theory*, vol. 58, no. 5, pp. 2541–2557, 2012.

100. P. Bühlmann and A. Wyner, “Variable length Markov chains,” *Ann. Stat.*, vol. 27, no. 2, pp. 480–513, 1999.

101. A. Berchtold and A. Raftery, “The mixture transition distribution model for high-order Markov chains and non-Gaussian time series,” *Statistical Science*, vol. 17, no. 3, pp. 328–356, 2002.

102. F. Willems, Y. Shtarkov, and T. Tjalkens, “Context weighting: General finite context sources,” in *14th Symposium on Information Theory in the Benelux*, Veldhoven, The Netherlands, May 1993.

103. ——, “Context tree weighting: Basic properties,” unpublished manuscript, summer 1993. Available online at:
www.sps.ele.tue.nl/members/F.M.J.Willems/RESEARCH_files/CTW/ResearchCTW.htm.

104. ——, “Context tree weighting: Basic properties,” *IEEE Trans. Inform. Theory*, vol. 41, no. 3, pp. 653–664, 1995.

105. ——, “Context tree maximizing,” in *2000 Conference on Information Sciences and Systems*, Princeton, NJ, March 2000.

106. ——, “Context tree weighting: Multi-alphabet sources,” in *14th Symposium on Information Theory in the Benelux*, Veldhoven, The Netherlands, May 1993.

107. F. Willems, A. Nowbahkt-Irani, and P. Volf, “Maximum a-posteriori tree models,” in *4th International ITG Conference on Source and Channel Coding*, Berlin, Germany, February 2002.

108. F. Willems and P. Volf, “Context maximizing: Finding MDL decision trees,” in *15th Symposium on Information Theory in the Benelux*, Louvain-la-Neuve, Belgium, May 1995.

109. ——, “A study of the context tree maximizing method,” in *16th Symposium on Information Theory in the Benelux*, Nieuwerkerk a/d IJssel, The Netherlands, May 1995.

110. F. Willems, “Coding for a binary independent piecewise-identically-distributed source,” *Information Theory, IEEE Transactions on*, vol. 42, no. 6, pp. 2210–2217, Nov 1996.

111. F. Willems, Y. Shtarkov, and T. Tjalkens, “Context weighting for general finite-context sources,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 1514–1520, 1996.

112. F. Willems, “The context-tree weighting method: Extensions,” *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 792–798, 1998.

113. A. Nowbakht and F. Willems, “Faster universal modeling for two source classes,” in *23rd Symposium on Information Theory in the Benelux*, Louvain-la-Neuve, Belgium, May 2002.

114. R. Begleiter, R. El-yaniv, and G. Yona, “On prediction using variable order Markov models,” *Journal of Artificial Intelligence Research*, vol. 22, pp. 385–421, 2004.

115. H. Cai, S. Kulkarni, and S. Verdú, “An algorithm for universal lossless compression with side information,” *Information Theory, IEEE Transactions on*, vol. 52, no. 9, pp. 4008–4016, Sept. 2006.

116. J. Ziv and N. Merhav, “On context-tree prediction of individual sequences,” *Information Theory, IEEE Transactions on*, vol. 53, no. 5, pp. 1860–1866, May 2007.

117. R. Gwadera, A. Gionis, and H. Mannila, “Optimal segmentation using tree models,” *Knowl. Inf. Syst.*, vol. 15, no. 3, pp. 259–283, May 2008.

118. C. Dimitrakakis, “Bayesian variable order Markov models,” in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. JMLR: W&CP, vol. 9, Chia Laguna Resort, Sardinia, Italy, 2010.

119. J. Veness, K. Ng, M. Hutter, and M. Bowling, “Context tree switching,” in *Data Compression Conference*, 2012, pp. 327–0336.

120. J. Rissanen, “A universal prior for integers and estimation by minimum description length,” *Ann. Statist.*, vol. 11, no. 2, pp. 416–431, 1983.

121. ——, “Complexity of strings in the class of Markov sources,” *Information Theory, IEEE Transactions on*, vol. 32, no. 4, pp. 526–532, Jul 1986.

122. M. Weinberger, J. Rissanen, and M. Feder, “A universal finite memory source,” *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 643–652, <ay 1995.

123. P. Bühlmann, “Model selection for variable length Markov chains and tuning the context algorithm,” *Annals of the Institute of Statistical Mathematics*, vol. 52, no. 2, pp. 287–315, 2000.

124. M. Mächler and P. Bühlmann, “Variable length Markov chains: Methodology, computing, and software,” *Journal of Computational and Graphical Statistics*, vol. 13, no. 2, pp. 435–455, 2004.

125. A. Garivier and F. Leonardi, “Context tree selection: A unifying view,” *Stochastic Processes and their Applications*, vol. 121, no. 11, pp. 2488–2506, 2011.

126. D. Ron, Y. Singer, and N. Tishby, “The power of amnesia: Learning probabilistic automata with variable memory length,” *Machine Learning*, vol. 25, pp. 117–149, 1996.

127. D. Dalevi and D. Dubhashi, “The Peres-Shields order estimator for fixed and variable length Markov models with applications to DNA sequence similarity,” in *Proceedings of the 5th International conference on Algorithms in Bioinformatics*, ser. WABI’05, 2005, pp. 291–302.

128. D. Duarte, A. Galves, and N. Garcia, “Markov approximation and consistent estimation of unbounded probabilistic suffix trees,” *Bulletin of the Brazilian Mathematical Society*, vol. 37, no. 4, pp. 581–592, December 2006.

129. I. Csiszar and Z. Talata, “Context tree estimation for not necessarily finite memory processes, via BIC and MDL,” *Information Theory, IEEE Transactions on*, vol. 52, no. 3, pp. 1007–1016, March 2006.

130. T. Roos and B. Yu, “Sparse Markov source estimation via transformed Lasso,” in *Networking and Information Theory, 2009. ITW 2009. IEEE Information Theory Workshop on*, June 2009, pp. 241–245.

131. J. Busch, P. Ferrari, A. Flesia, R. Fraiman, S. Grynberg, and F. Leonardi, “Testing statistical hypothesis on random trees and applications to the protein classification problem,” *Annals of Applied Statistics*, vol. 3, no. 2, pp. 542–563, 2009.

132. F. Leonardi, “Some upper bounds for the rate of convergence of penalized likelihood context tree estimators,” *Brazilian Journal of Probability and Statistics*, vol. 24, no. 2, pp. 321–336, 2010.

133. A. Abakuks, “The synoptic problem: On Matthew’s and Luke’s use of Mark,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 175, no. 4, pp. 959–975, 2012.

134. A. Galves, C. Galves, J. Garcia, N. Garcia, and F. Leonardi, “Context tree selection and linguistic rhythm retrieval from written texts,” *Annals of Applied Statistics*, vol. 6, no. 1, pp. 186–209, 2012.

135. N. Kalouptsidis, *Signal Processing Systems Theory and Design*. John Wiley and Sons, 1997.

136. S. Boyd, “Volterra series: Engineering fundamentals,” Ph.D. dissertation, UC Berkeley, 1985.

137. S. Boyd and L. Chua, “Fading memory and the problem of approximating nonlinear operators with Volterra series,” *IEEE Transactions on Circuits and Systems*, vol. 32, no. 11, pp. 1150–1161, Nov. 1985.

138. J. Brewer, “Kronecker products and matrix calculus in system theory,” *IEEE Transactions on Circuits and Systems*, vol. 25, no. 9, pp. 772 – 781, sep 1978.

139. S. Benedetto and S. Biglieri, *Principles of Digital Transmission: with wireless applications*. Kluwer Academic, 1998.

140. F. Doyle, R. Pearson, and B. Ogunnaike, *Identification and control using Volterra series*. Springer, 2002.

141. D. Westwick and R. Kearney, *Identification of nonlinear physiological systems*. IEEE Press, 2003.

142. G. Giannakis and E. Serpedin, “Linear multichannel blind equalizers of nonlinear FIR Volterra channels,” *IEEE Transactions on Signal Processing*, vol. 45, no. 1, pp. 67 –81, jan 1997.

143. F. Gregorio, J. Cousseau, S. Werner, T. Riihonen, and R. Wichman, “Power amplifier linearization technique with iq imbalance and crosstalk compensation for broadband MIMO-OFDM transmitters,” *EURASIP Journal on Advances in Signal Processing*, 2011.

144. C. A. R. Fernandes, “Nonlinear MIMO communication systems: Channel estimation and information recovery using Volterra models,” Ph.D. dissertation, Universite de Nice-Sophia Antipolis, 2009.

145. F. Gregorio, S. Werner, T. I. Laakso, and J. Cousseau, “Receiver cancellation technique for nonlinear power amplifier distortion in SDMA–OFDM systems,” *IEEE Transactions on Vehicular Technology*, vol. 56, no. 5, pp. 2499 –2516, Sept. 2007.

146. A. J. Redfern and G. T. Zhou, “Blind zero forcing equalization of multichannel nonlinear CDMA systems,” *IEEE Transactions on Signal Processing*, vol. 49, no. 10, p. 2363–2371, Oct. 2001.

147. D. Angelosante, E. Grossi, G. Giannakis, and M. Lops, “Sparsity–aware estimation of CDMA system parameters,” *EURASIP Journal on Advances in Signal Processing*, 2010.