MINLIP for the Identification of Monotone Wiener Systems ⋆

Kristiaan Pelckmans [a]

[a] *Division of Systems and Control, Department of Information Technology*
Uppsala University, Box 337, SE-751 05, Uppsala, Sweden

Abstract

This paper studies the MINLIP estimator for the identification of Wiener systems consisting of a sequence of a linear FIR dynamical model, and a monotonically increasing (or decreasing) static function. Given T observations, this algorithm boils down to solving a convex quadratic program with $O(T)$ variables and inequality constraints, implementing an inference technique which is based entirely on model complexity control [1] . The resulting estimates of the linear submodel are found to be almost consistent when no noise is present in the data, under a condition of smoothness of the true nonlinearity and local Persistency of Excitation (local PE) of the data. This result is novel as it does not rely on classical tools as a 'linearization' using a Taylor decomposition, nor exploits stochastic properties of the data. It is indicated how to extend the method to cope with noisy data, and empirical evidence is providing contrasting the estimator against other recently proposed techniques.

Key words: System Identification, Wiener Systems, Nonlinear Systems, Quantized signals, Convex Optimization

1 INTRODUCTION

The identification of Wiener systems has been considered in many papers since the 1970s. Different existing approaches could roughly be divided in methods using (i) Invertible nonlinearities (reducing to Hammerstein identification), (ii) correlation based approaches exploiting stochastic properties of the signals [3,1], (iii) approximate (recursive) PEM approaches providing a well-established framework for convergence analysis [16,17] and [6], (iv) subspace based approaches [15]. For a general overview see the survey [5]. Specific applications towards identification with quantized outputs are considered in [18], see also the book [12]. The present approach builds further on ideas developed in [19,2]. Another relevant work is [14] who consider a similar identification task as we will do. The MINLIP estimator (or shortly MINLIP, we will carify the abbrevaiation shortly) studied in this paper for identification of dynamic Wiener systems was originally conceived in the context of learning ranking functions and survival analysis, see [13] and earlier work of those authors. In [9], the authors studied the impact of model complexity control for the identification of Hammerstein systems. In [10] the use of explicit complexity control was investigated in the context of adaptive filtering.

While the literature on the identification of Wiener systems is considerable, often theoretical understanding of the proposed techniques is restricted to exposition of an appropriate technical implementation. Notable exceptions are given in [17], [5] and [2]. The first work considers a Recursive Prediction Error Method (RPEM) of general Wiener models, and convergence properties are derived using the ODE framework as in [7]. This approach makes considerable assumptions on the stochastic mechanisms underlying the signals used for (recursive) identification. The smoothing approach described in [5] exploits as well a stochastic assumption of the involved signals, and asserts basically that the Wiener system can be identified by directly averaging out the nonlinear effect. Although powerful concentration inequalities lie on the basis of this approach, no argument is given that this method applies for Wiener systems which are more complex (realistic) than the academic examples presented in those papers.

This work was prompted by the earlier [19], exploring the task of identification of monotone wiener systems. The practical algorithm proposed in that paper will always yield a trivial solutions ($h = 0_d$ in their notation), and is as such to be depreciated. The line of thinking however looks powerful, and this led [2] to investigate the question under what conditions on the static nonlinearity (besides monotonicity) a Wiener system is identifiable. The present work takes this results a step further, introducing model complexity control

⋆ The author acknowledges financial support of European Research Council under Seventh Framework Program and within Advanced Grant no. 247035 "Systems and Signals Tools for Estimation and Analysis of Mathematical Models in Endocrinology and Neurology".

Email address: kp@it.uu.se (Kristiaan Pelckmans).

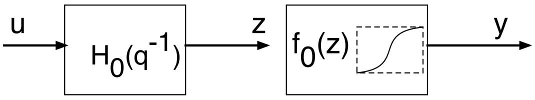

Fig. 1. *Schematical representation of Wiener systems under consideration. The function $f : \mathbb{R} \to \mathbb{R}$ is assumed to be monotonically increasing. Neither $H(q)$ nor f is assumed to be invertible.*

into the picture. Specifically, we express model complexity in terms of a Lipschitz property of the estimated nonlinearity, The idea of minimizing (MIN-) this Lipschitz property (-LIP) results directly in an efficient identification technique termed the MINLIP estimator - or shortly MINLIP - which can be solved efficiently using tools of convex optimization. Specifically, we rephrase the identification task as a convex Quadratic Programming (QP) problem of $O(T)$ unknowns and inequalities (where T denotes the number of samples). We as well present an analysis that the estimates given by this algorithm are almost consistent, where the approximation factor relies on the *richness* of the data in terms of a local measure of persistency of excitation, and the smoothness of the static nonlinearity. This analysis does not resort to local approximations using a Taylor decomposition, and does not need any stochastic setup.

The contribution of this paper is threefold. Section 2 describes the precise class of monotone Wiener systems which is envisaged, and discusses the main ideas motivating the MINLIP estimator. An artificial yet challenging case study provides empirical evidence for this estimator for noiseless data. Section 3 then establishes almost consistency of the estimates under appropriate conditions of the data used for identification, and the underlying model. This result is nontrivial as the considered model does not allow for a straightforward (finite) parametrization, and is not based on minimizing a model mismatch criterion as classical. In particular, we need to have that the data is locally Persistent Exciting (PE) while the true monotone nonlinearity needs to be smooth around its steepest part.

Section 3 describes an extension towards the case where (i) data is perturbed by noise, (ii) the true system does not belong to the considered model class of monotone Wiener systems of given order, or (iii) where no true model is assumed to exist. Again, the MINLIP estimates are given by solving a convex Quadratic Program with $O(T)$ unknowns and inequality constraints. Empirical evidence is given for the use of this estimator, and the degradation of the accuracy of the estimate in case of noise is investigated experimentally. Section 5 concludes the paper, and highlights a number of open research questions.

2 Identification of Monotone Wiener Systems

2.1 The Model Class

This work focus on the identification of nonlinear dynamic models in the following model class.

Definition 1 (FIR Wiener Model (f, a)) *A FIR Wiener model consists of a sequence of (i) a linear dynamical model characterized by an impulse response function $H(q^{-1})$ (here q^{-1} is the backshift operator as classically) applied on the* input *signal $\{u_t\}_t$, and (ii) a static nonlinear function $f : \mathbb{R} \to \mathbb{R}$ (see Fig. 1). If the signals $\{u_t\}_t$ and $\{y_t\}_t$ follow such a model with 'true' subsystem H_0 and 'true' function f_0, we can write*

$$y_t = f_0\Big(H_0(q^{-1})(u_t)\Big), \tag{1}$$

and we say that the observations come from the Wiener system (H_0, f_0). *For a FIR-Wiener model of order d, one considers a Finite Impulse Response (FIR) parametrization of the linear subsystem, or $H(q^{-1}) = a_1 q^{-1} + \cdots + a_d q^{-d}$. We denote such model (in the context of this paper) shortly as the (f, a)-Wiener model. Now since a and the domain of f can be rescaled arbitrarily, it is convenient to impose $\|a\|_2 = 1$, which does avoid identifiabilty issues. If the signals $\{u_t\}_t$ and $\{y_t\}_t$ obey such a model with 'true' function f_0 and 'true' parameters a_0, or*

$$y_t = f_0\left(\sum_{k=1}^{d} a_{0,k} u_{t-k}\right) = f_0(a_0^T \mathbf{u}_t). \tag{2}$$

- where $a_0 \in \mathbb{R}^d$ and $\|a_0\|_2 = 1$ and we define $\mathbf{u}_t = (u_{t-1}, \ldots, u_{t-d})^T \in \mathbb{R}^d$ - we say that the observations come from the Wiener system (a_0, f_0). *We will denote the set of possible observations as $\mathcal{S} = \{(\mathbf{u}_t, y_t)\}_t \subseteq \mathbb{R}^d \times \mathbb{R}$.*

We specialize further to a subset of this class as follows, schematically illustrated in fig. (1).

Definition 2 (Monotone FIR Wiener Model (f, a)) *A FIR-Wiener model (f, a) is called* monotone *if $f : \mathbb{R} \to \mathbb{R}$ is monotonically increasing (but not necessarily invertible),*

$$y_t = f(a^T \mathbf{u}_t). \tag{3}$$

We define the Monotone Wiener model class formally as

$$\mathcal{F} = \Big\{(f, a) \;\Big|\; f : \mathbb{R} \to \mathbb{R} : \textit{Monotonically increasing}, \\ a \in \mathbb{R}^d, \|a\|_2 = 1\Big\}. \tag{4}$$

Note that by similarity one has $(f, a) = (f', -a)$, where $f(z) = f'(-z)$ for all $z \in \mathbb{R}$. Now f' is monotonically decreasing, explaining why we can omit the denominator 'increasing' in the nomenclature. As argued before, this class of monotone FIR-Wiener model can capture such different effects as (1) quantized output measurements, (2) saturation effects of the sensor, and (3) handling of general bijective transformations of the output scaling (cfr. the temperature scale of Celsius versus Fahrenheit), amongst others.

2.2 Identification by MINLIP

The identification technique implements the adagio 'make problems as simple as possible, but not simpler'. The surprising result is that this idea may yield consistent estimates, without any reference to notions as 'statistical likelihood' or 'prediction error'.

The problem of identification of a Wiener system from observations is traditionally formalized as follows

$$\min_{a,f:\ \|a\|_2=1} J(f,a) = \sum_{t=d+1}^{T} \left(f(a^T \mathbf{u}_t) - y_t\right)^2. \qquad (5)$$

We refer to this formulation of the estimation problem as to a prediction error method for Wiener models - abbreviated here as WPEM - and it will be mainly this approach we will contrast the proposed method against. Note that this approach in case of noise as in Section III is a merely an Output-Error (OE) technique, unless stringent stochastic assumptions can be made on the noise, see e.g. [6] for a discussion. In general,this formulation is hard to solve as the unknowns a interact directly with the unknown function f. As a result one typically resorts to an iterative scheme or general purpose nonlinear optimization routine. The practical procedures lack generality and robustness for different reasons (i) depending on the form (or parametrization) of f, ill-conditioning of the problem may arise or even gradient information may not exist, (ii) the problem can often be stuck in local minima, which can be arbitrary bad (iii) procedures are highly depending on the exact representation of the unknown f. In general, such procedures are therefore not easily scalable to more complex settings.

The approach we will advocate in this paper is however conceptually quite different. *Rather than minimizing the equation errors, we look for the least complex model reconstructing the observations*. What we mean by 'least complex model' is somehow up to the user to decide. In this paper we consider a specific complexity measure defined as follows which will eventually reduce the inference problem to an optimization problem which can be solved efficiently, and for which we prove consistency.

Definition 3 (Lipshitz Condition) *Consider a functionf : $\mathbb{R} \to \mathbb{R}$. Assume there exists a constant L such that one has for all $z, z' \in \mathbb{R}^d$ that*

$$|f(z) - f(z')| \leq L\,|z - z'|, \qquad (6)$$

then f is Lipschitz smooth with a constant L.

Now, we structure the model class by the following nested sets

$$\mathcal{F}_L = \left\{(f,a) \in \mathcal{F} \;\middle|\; f: \text{Lipschitz with constant } L, \|a\|_2 = 1\right\}. \qquad (7)$$

Now if $L_1 < L_2 < \cdots < L_k$ are k sorted constants, then one has a nested structure over the model class, i.e.

$$\mathcal{F}_{L_1} \subseteq \mathcal{F}_{L_2} \subseteq \cdots \subseteq \mathcal{F}_{L_k} \subseteq \mathcal{F}. \qquad (8)$$

A plausible identification algorithm would now be *to find the linear parameters $a \in \mathbb{R}^d$ such that the mapping from $\{a^T \mathbf{u}_t\}_t$ to the corresponding values $\{y_t\}_t$ has as small a Lipschitz value L as possible*. Since we are only interested at this stage in the parameters a rather than also recovering f, we focus attention on the given samples only. Consequently, sufficient condition for our needs for a function f to be Lipschitz smooth with constant L is

$$\left|f(a^T\mathbf{u}_i) - f(a^T\mathbf{u}_j)\right| \leq L\left|a^T(\mathbf{u}_i - \mathbf{u}_j)\right|, \forall i < j = d, \ldots, T. \qquad (9)$$

By exploiting the monotonicity property of f, one can write the $O(T^2)$ constraints equivalently using only $O(T)$ constraints by ordering the data. This step captures the function f implicitly (see Figure (2)), proven as follows.

Lemma 1 (Existence of a Transformation Function) *Given a collection of pairs $\{(z_{(i)}, y_{(i)})\}_{i=1}^n$, enumerated such that $y_{(i)} \leq y_{(j)}$ if and only if $i \leq j$. Then we consider the sample conditions for $L < \infty$:*

$$0 \leq \left(y_{(j)} - y_{(i)}\right) \leq L\left(z_{(j)} - z_{(i)}\right) \quad \forall i < j = 1, \ldots, n, \qquad (10)$$

(1) If one has for a finite $L > 0$ that (10) holds, then there exist a monotonically increasing function $f : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant L interpolating the samples.

(2) If one has that for all admissible $(z, y) \in \mathbb{R} \times \mathbb{R}$ one has that $y = f(z)$ for an (unknown) continuous, (finite) differentiable and monotonically increasing function $f : \mathbb{R} \to \mathbb{R}$, then there is an $L < \infty$ such that (10) holds.

Proof: To proof 1, consider the linear interpolation function $f_n : \mathbb{R} \to \mathbb{R}$, defined as

$$f_n(z) = \frac{z - z_{\underline{z}(z)}}{z_{\overline{z}(z)} - z_{\underline{z}(z)}}\left(y_{\overline{z}(z)} - y_{\underline{z}(z)}\right) + y_{\underline{z}(z)}, \qquad (11)$$

where we define $\overline{z}(z) = \arg\min_i(z_i : z_i \geq z)$ and $\underline{z}(z) = \arg\max_i(z_i : z_i \leq z)$. Direct manipulation shows that this function is monotonically increasing and continuous. Now take $z < z' \in \mathbb{R}$, then we have to show that $f_n(z') - f_n(z) \leq L(z' - z)$. For convenience of notation define $l = \underline{z}(z)$, $u = \overline{z}(z)$, $l' = \underline{z}(z')$ and $u' = \overline{z}(z')$, then

$$\begin{aligned}\frac{z' - z_{l'}}{z_{u'} - z_{l'}}(y_{u'} - y_{l'}) - \frac{z - z_l}{z_u - z_l}(y_u - y_l) + (y_{l'} - y_l)\\ \leq L(z' - z_{l'}) - L(z - z_l) + L(z_{l'} - z_l)\\ = L(z' - z), \qquad (12)\end{aligned}$$

since $z_l \leq z_{l'}$ and $y_l \leq y_{l'}$ by definition.

Item 2 is proven as follows. Let f' be the derivative of a differentiable function f_0, then the mean value theorem asserts that for any two samples (z_i, y_i) and (z_j, y_j) for which $z_i \leq z_j$, there exists a $z \in (z_i, z_j) \subset \mathbb{R}$ such that

$$(y_j - y_i) = (z_j - z_i) f'(z) \leq L(z_i - z_j) \tag{13}$$

where $L = \sup_z f'(z)$. This observation motivates the following procedure: find parameters a such that f has minimal Lipschitz condition. The solution is given by solving

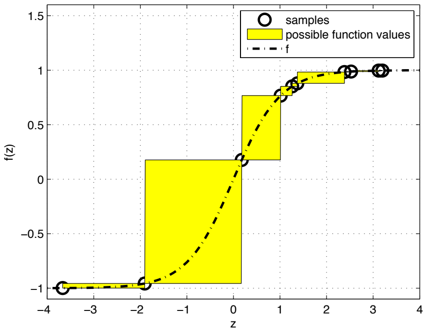

Fig. 2. *Schematic representation of Lemma 1. If a Lipschitz-smooth monotone f exists (black dash-dotted curved line), samples obey the pairwise Lipschitz constraint. If samples exist satisfying the Lipschitz constraints, one can always find a monotone function interpolating this samples (indicated by the yellow blocks).*

$$\max_a \min_{y_i \neq y_j} \frac{|a^T(\mathbf{u}_i - \mathbf{u}_j)|}{|y_i - y_j|}$$
$$\text{s.t.} \quad a^T \mathbf{u}_{(i)} \geq a^T \mathbf{u}_{(i-1)}, \quad \forall i = d+1, \ldots, T, \tag{14}$$

or

$$\min_{a,L} \; L^2 \quad \text{s.t.} \quad \|a\|_2 = 1,$$
$$(y_{(i)} - y_{(i-1)}) \leq L \left(a^T (\mathbf{u}_{(i)} - \mathbf{u}_{(i-1)}) \right), \; \forall i = d+1, \ldots, T. \tag{15}$$

where we have only a linear number of constraints.After a change of variable, we can write equivalently

Definition 4 (MINLIP) *Given an ordered set of examples $\{(\mathbf{u}_{(i)}, y_{(i)})\}_{i=d}^T \subset \mathbb{R}^d \times \mathbb{R}$ indexed such that $y_{(i-1)} \leq y_{(i)}$ for all $i = d+1, \ldots, T$, then our (rescaled) estimate $\hat{a}$ follows by solving*

$$a_T = \arg\min_a \; a^T a \;\; \textit{s.t.}$$
$$(y_{(i)} - y_{(i-1)}) \leq a^T (\mathbf{u}_{(i)} - \mathbf{u}_{(i-1)}), \; \forall i = d+1, \ldots, T. \tag{16}$$

where the estimated function $\hat{f}$ is specified only implicitly as in Proposition 1.

This problem can be cast as a convex Quadratic Program (QP) with $T - d$ linear constraints and d unknowns. This problem can be solved efficiently with contemporarily solvers available in most mathematical packages [2] . The a_T which minimizes this constrained objective is our estimate of a (rescaled) version of the parameters of the FIR system $H(q)$. The problem is written in matrix notation as

$$a_T = \arg\min_{a \in \mathbb{R}^d} a^T a \text{ s.t. } (\Delta \mathbf{u}) a \leq \Delta y, \tag{17}$$

where $\mathbf{u} = (\mathbf{u}_{(d)}, \ldots, \mathbf{u}_{(T)}) \in \mathbb{R}^{(T-d+1) \times d}$ is a Hankel matrix (up to the sorting!), $y = (y_{(d)}, \ldots, y_{(T)}) \in \mathbb{R}^{T-d+1}$ is an ordered vector and

$$\Delta = \begin{bmatrix} -1 & 1 & 0 & & 0 \\ 0 & -1 & 1 & & \\ & & \ddots & \ddots & \\ 0 & & & -1 & 1 \end{bmatrix} \in \{-1, 0, 1\}^{(T-d) \times (T-d+1)}. \tag{18}$$

In order to improve reproducibility of the result and stress practical use, a full MATLAB implementation in 8 lines of code is given in Alg. (1). In order to extend this implementation to handle general cases one should take additional care of (possible) ties of output samples (See Subsection IV.C for a practical way to cope with this issue).

Algorithm 1 A MATLAB implementation of MINLIP

```
n  = length(y);
x1 = toeplitz(u(d:end),u(d:-1:1));
y1 = y(d:end);
[ys,si]=sort(y1);
xs = x1(si,:);
e  = ones(n,1);
D  = full(spdiags([-e e],[0 1],n-d,n-d+1));
a  = quadprog(eye(d),zeros(d,1),-D*xs,-D*ys);
```

In a second phase, it could be useful to reconstruct f_0 based on a_T and the samples $\{(a_T^T \mathbf{u}_t, y_t)\}_{t=d}^T$. We suggest to use the linear interpolation defined in (12) to proof Lemma 1. This function is by construction Lipschitz smooth with constant $L = \sqrt{a_T^T a_T}$. It is not too difficult to come up with more parsimonious estimators of the univariate function f_0 based on the bivariate samples $\{(a_T^T \mathbf{u}_t, y_t)\}_{t=d}^T$ which behaves more robust against modeling errors.

2.3 Almost Consistency of MINLIP in the Noiseless Case

This section characterizes how well the estimate approach the true impulse response, under suitable assumptions on the data and the static nonlinearity. Particularly, we assume that the observations arise from a *true* Monotone Wiener model with FIR system of given order for the dynamic part, and

[2] In our experiments we use the solver available at http://www.mosek.org.

that no noise perturbs the observations. This problem is nontrivial as (i) the proposed model is essentially nonlinear and non-parametric, and (ii) the method is not based directly on minimizing a mismatch between the model and the observations. The main outcome is that (approximate) consistent estimates are given when the data satisfies a condition of local Persistency of Excitation (PE), and the true monotone static nonlinearity is smooth around its steepest part.

This result is referred to here as *almost consistency*, as it guarantees accuracy of the estimates only up to a small (but often non-zero) approximation term. In a sense, this is the best one could hope for here because of two reasons,: (a) the model is non-parametric (or semi-parametric) as the static monotone function of the model cannot be expressed straightforwardly in terms of a (small number of) parameters. In that respect, a finite dataset contains never have enough information in order to reconstruct this system exactly. (b) A finite dataset can never be locally exciting in every (arbitrary small) neighborhood, but can only guarantee this condition for all localities which are sufficiently large. Classical concepts as bias or variance do not cover this notion as no stochastic assumptions are made. The question wether approximate consistency implies asymptotic consistency requires one as well to make additional stochastic assumptions underlying the data, and is not covered in this text as such. The analogue for linear estimating of a FIR model goes as follows: assume the system can be described exactly as a FIR model of given order (smaller than) d, then the corresponding notion is that the least squares estimate is *exactly* consistent if the data is (globally) PE to an order d. Since we assume that there is no noise in the data, there is no approximation to be made here. This difference can be seen as the consequence of the semi-parametric model of the Monotone Wiener system where in general no finite/small parameterization exists matching the system.

Assume that the observed system obeys the relation given in (32) with a fixed (but unknown) monotonically increasing function $f_0 : \mathbb{R} \to \mathbb{R}$ which is Lipschitz monotone with constant $L_0 < \infty$, and parameter vector $a_0 \in \mathbb{R}^d$. We refer to those as to the *true function* f_0 and the *true parameters* a_0 respectively. We address the question wether if we see enough data (or $T \to \infty$), the MINLIP estimate a_T will equal a_0 up to a scaling constant. Formally, we consider the MINLIP estimator based on the (infinite) set $\mathcal{S}$ as

$$\begin{aligned} &\hat{L} = \min_{L, \|a\|_2=1} L \\ &\text{s.t. } (y-y') \le L a^T(\mathbf{u}-\mathbf{u}'),\ \forall (\mathbf{u}, y), (\mathbf{u}', y') \in \mathcal{S}, y > y'. \end{aligned} \tag{19}$$

Sometimes it will be convenient to rewrite the MINLIP estimator (19) as the following minimax problem:

$$\ell = \max_{\|a\|_2=1} \inf_{(\mathbf{u},y),(\mathbf{u}',y)\in\mathcal{S}: y>y'} \frac{a^T(\mathbf{u}-\mathbf{u}')}{y-y'}. \tag{20}$$

where $\ell \ge \frac{1}{L_0}$ by construction of L_0. In order to characterize the solution, the following two conditions are needed.

Definition 5 (f is (L_0, g)-Lipschitz on $\mathcal{S}' \subseteq \mathbb{R}$) *The function f is said to be (L_0, g)-Lipschitz on $\mathcal{S}' \subseteq \mathbb{R}$ for a decreasing, positive function $g : \mathbb{R}^+ \to \mathbb{R}^+$ with $g(0) = 1$ if: (A) one has for all $z, z' \in \mathcal{S}'$ that*

$$(f(z) - f(z')) \le L_0 (z - z'). \tag{21}$$

(B) there exists a $z \in \mathcal{S}'$ and $z' \in \mathcal{S}'$ such that

$$(f(z) - f(z')) = L_0 (z - z'), \tag{22}$$

and (C) one has for this z, for any $\epsilon > 0$ and $z'' \in \mathcal{S}'$ where $|z - z''| \le \epsilon$ that

$$|f(z) - f(z'')| \ge g(|z - z''|) L_0 |z - z''|. \tag{23}$$

Hence g denotes how 'smooth' the constant L decays in a neighborhood of z where the actual Lipschitz constraint is met (that is, a slower decaying function g indicates a higher smoothness). In particular, a value $h(\epsilon) = 1$ implies that the function f is linear with slope L_0 in this neighborhood. Such characterization is illustrated for $f(z) = \tanh(z)$ in Figure (3) for a smoothness function $g(\epsilon) = 1/(1 + c\epsilon)$.

Definition 6 (ϵ-Local Persistently Exciting) *We say that a set $\mathcal{S} \subseteq \mathbb{R}^d$ is ϵ-local persistent exciting of order d for $\epsilon > 0$ iff for any vector $\mathbf{u} \in \mathcal{S}$, there exist d vectors $\mathbf{u}_1, \ldots, \mathbf{u}_d \in \mathcal{S}$ with $\{\mathbf{u} - \mathbf{u}_k\}_{k=1}^m$ linearly independent vectors and*

$$\|\mathbf{u} - \mathbf{u}_k\|_2 \le \epsilon,\ \forall k = 1, \ldots, d. \tag{24}$$

This definition can be seen as a *local* version of Persistency of Excitation (PE), see e.g. [8,11,4] for the classical definition of PE.

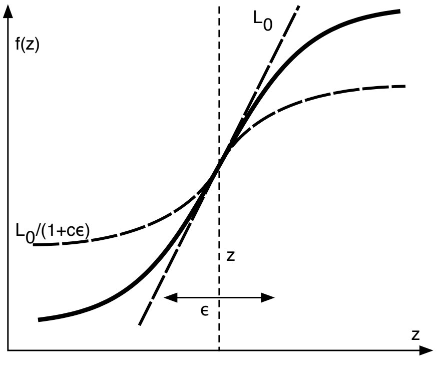

Fig. 3. *Schematic illustration of the (L_0, g)-Lipschitz property of a function f with $g(\epsilon) = \frac{1}{1+c\epsilon}$. There should be a sample z where the Lipschitz constant L_0 is attained, and in the ϵ-neighborhood of this sample z the Lipschitz-property shouldn't decay too fast, e.g. in the neighborhood of z the function behave almost linearly.*

Theorem 1 (Almost Consistency) *Fix $\epsilon > 0$ and consider the (f_0, a_0)-Wiener system as in (3) with corresponding observations in $\mathcal{S}$. If $f_0 : \mathbb{R} \rightarrow \mathbb{R}$ is (L_0, g)-Lipschitz and monotone on the set $\{(z, y) : z = a_0^T \mathbf{u} : \mathbf{u} \in \mathcal{S}\}$ and $\mathcal{S}$ is ϵ-local PE, then*

$$a_T^T a_0 \geq g(\epsilon), \tag{25}$$

where a_T is the estimate of MINLIP as in (16).

This means that the smoother the function f_0 is towards its steepest part, the better estimates we get. In particular, when f_0 is (almost) linear - we only need global PE to have exact estimates $a_T \propto a_0$. Specifically, if $g(\epsilon) = 1$ - meaning that the function h is linear - only global persistency of excitation is required to have consistency, and the MINLIP estimator will return the same result as a linear estimator.

Proof: Let the Lipschitz constant be achieved in the sample $(\mathbf{u}, y), (\mathbf{u}', y') \in \mathcal{S}$, such that

$$(y - y') = L_0(\mathbf{u} - \mathbf{u}')^T a_0. \tag{26}$$

As the set $\mathcal{S}$ is ϵ-local persistent exciting of order d, one can find for the sample $(\mathbf{u}, y) \in \mathcal{S}$ d vectors $\mathbf{u}_1, \ldots, \mathbf{u}_d$ contained in $\mathcal{S}$ such that the vectors $(\mathbf{u} - \mathbf{u}_1), \ldots, (\mathbf{u} - \mathbf{u}_d)$ are linearly independent and have norm smaller than ϵ. This implies that one can rewrite $a_0, a_T \in \mathbb{R}^d$ (i.e. the true parameter vector and the optimal estimate associated to (19)) as

$$\begin{cases} a_0 = \sum_{k=1}^d \alpha_{0,k} \overline{(\mathbf{u} - \mathbf{u}_k)} \\ a_T = \sum_{k=1}^d \alpha_k \overline{(\mathbf{u} - \mathbf{u}_k)}, \end{cases} \tag{27}$$

where $\alpha, \alpha_0 \in \mathbb{R}^d$, and we let $(\mathbf{u} - \mathbf{u}_k) = \sigma_k \overline{(\mathbf{u} - \mathbf{u}_k)} \|(\mathbf{u} - \mathbf{u}_k)\|_2$ with $\sigma_k \in \{-1, 1\}$ such that $\|\overline{(\mathbf{u} - \mathbf{u}_k)}\|_2 = 1$ and $\overline{(\mathbf{u} - \mathbf{u}_k)}^T a_0 \geq 0$ for all $k = 1, \ldots, d$. Define as previously for each $k = 1, \ldots, d$ the constant $L_k \in \mathbb{R}_+$ such that $(y - y_k) = L_k(\mathbf{u} - \mathbf{u}_k)^T a_0$, where $L_k \leq L_0$ by construction. Now define the matrix $D_\mathbf{u} \in \mathbb{R}^{d \times d}$ as

$$D_\mathbf{u} = \begin{bmatrix} \overline{(\mathbf{u} - \mathbf{u}_1)} \\ \overline{(\mathbf{u} - \mathbf{u}_2)} \\ \vdots \\ \overline{(\mathbf{u} - \mathbf{u}_d)} \end{bmatrix}, \tag{28}$$

such that $D_\mathbf{u} a_0 \geq 0_d$, and the matrix $L \in \mathbb{R}^{d \times d}$ as $L = \operatorname{diag}(L_1, \ldots, L_d)$. Then we have in matrix notation that $a_0 = D_\mathbf{u}^T \alpha_0$ and $a_T = D_\mathbf{u}^T \alpha$. By construction of the MINLIP estimator we have that

$$L D_\mathbf{u} D_\mathbf{u}^T \alpha_0 = L D_\mathbf{u} a_0 \leq \ell D_\mathbf{u} a_T = \ell D_\mathbf{u} D_\mathbf{u}^T \alpha, \tag{29}$$

where ℓ is the minimal value obtained in (19). As such $\frac{1}{\ell} L D_\mathbf{u} D_\mathbf{u}^T \alpha_0 \leq D_\mathbf{u} D_\mathbf{u}^T \alpha$. Then one has

$$\begin{aligned} a_0^T a_T = \alpha_0^T (D_\mathbf{u} D_\mathbf{u}^T) \alpha &\geq \frac{1}{\ell} \alpha_0^T L (D_\mathbf{u} D_\mathbf{u}^T) \alpha_0 \\ &\geq \frac{g(\epsilon) L_0}{\ell} \alpha_0^T (D_\mathbf{u} D_\mathbf{u}^T) \alpha_0 \\ &\geq g(\epsilon) \alpha_0^T (D_\mathbf{u} D_\mathbf{u}^T) \alpha_0 \geq g(\epsilon), \end{aligned} \tag{30}$$

since $\ell \leq L_0$. This proofs the result.

3 IDENTIFICATION WITH NOISY DATA

This section considers how to modify the estimator towards the case of noise being present in the data, or where the data-samples are only approximated by a monotone FIR (f, a)-system. Empirical evidence is provided for the use of this estimator.

3.1 Model Class

There are a number of different ways one can model noise in the class of monotone wiener systems. A first one is to consider noise on the measured outputs (or *measurement noise*). One may argue that this model is not a very realistic assumption in case the observations are a quantized version of the output of the linear system. That is, once the signal is quantized (and transmitted), it can often be measured without error. On the other hand, it is often not clear which noise model of a quantized signal (with a finite number of different levels) fits the application (a Gaussian distribution would not make much sense here). Another assumption one can make is that noise occurs in the signal $\{u_t\}_t$, but as this results in coloring of the noise by the unknown linear subsystem, this model is not adopted as yet. A third alternative is that the noise comes in between the linear subsystem and the monotone static function (see Fig. 4). As in the following no restrictive assumptions (as whiteness of the noise signal) is assumed, this could be seen as uncertainty coming in in the model by under-modelling of the linear system. This is the view which underlies the following definitions.

Definition 7 (Noisy FIR Wiener Model (f, a)) *A FIR Wiener model consists of a sequence of (i) a linear dynamical model characterized by an impulse response function $H(q^{-1})$ applied on the* input *signal $\{u_t\}_t$, (ii) a static non-linear function $f : \mathbb{R} \rightarrow \mathbb{R}$, and (iii) a sequence of 'noise' terms $\{e_t\}_t$. If the signals $\{u_t\}_t$, $\{y_t\}_t$ and $\{e_t\}_t$ follow such a model with 'true' subsystem H_0 and 'true' function f_0, we can write*

$$y_t = f_0\Big(H_0(q^{-1})(u_t) + e_t\Big), \tag{31}$$

If the signals $\{u_t\}_t$, $\{y_t\}_t$ and $\{e_t\}_t$ obey a Wiener FIR-

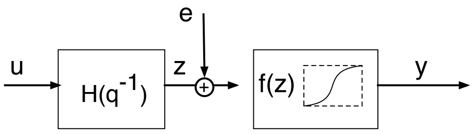

Fig. 4. *Schematic representation of a noisy monotone wiener system. This paper adopts the setting that noise comes in after the linear dynamic part (capturing model mismatch), and right before application of the static nonlinearity (or quantization).*

model with 'true' function f_0 and 'true' parameters a_0, or

$$y_t = f_0\left(\sum_{k=1}^{d} a_{0,k}u_{t-k} + e_t\right) = f_0(a_0^T\mathbf{u}_t + e_t), \quad (32)$$

where $a_0 \in \mathbb{R}^d$ and $\|a_0\|_2 = 1$.

3.2 MINLIP for Noisy Data

Given time-series $\{u_t\}_t$ and $\{y_t\}_t$, referred to as 'input' and 'output'. Again, let $\{(\mathbf{u}_t = (u_{t-d+1},\ldots,u_t), y_t)\}_{t=d}^T \subset \mathbb{R}^d \times \mathbb{R}$ be a dataset containing $T-d+1$ samples. Let this set be reindexed as $\{(\mathbf{u}_{(j)}, y_{(j)})\}_{j=d}^T$ where $y_{(i)} \leq y_{(j)}$ for all $d < i < j \leq T$. Then adopting the noisy model (32) suggests modification of the standard MINLIP (see eq. (16)) as

$$\begin{aligned} &\min_{a,e}\ \frac{1}{2}a^Ta + \frac{\gamma}{2}\sum_{t=d}^{T}|e_i| \\ &\text{s.t.}\quad (y_{(i)} - y_{(i-1)}) \leq (a^T\mathbf{u}_{(i)} + e_{(i)}) - (a^T\mathbf{u}_{(i-1)} + e_{(i-1)}), \\ &\qquad\qquad \forall i = d+1,\ldots,T, \quad (33)\end{aligned}$$

where the fixed *regularization parameter* $\gamma > 0$ trades the Lipschitz based regularization term and the penalization of the residuals. The choice of penalizing the absolute loss of the residuals is inspired by (i) robustness considerations and the (ii) non-stochastic nature of the residuals where a worst-case approach is more suited. The tuning of this constant can be done with an appropriate model selection criterion as cross-validation. As before, this optimization problem can be solved efficiently as a convex quadratic program (QP) using $O(T)$ unknowns and inequality constraints.

4 Empirical Evidence

This section spells out a number of artificial yet challenging case studies. A main argument which is made here is that although the underlying system H_0 may be represented as a fractional polynomial, it is often useful to consider a FIR model consisting of a large number of tapped delays (say $d = O(100)$) which can approach (the impulse response function of) H_0 arbitrarily close. In this way, one does not have to specify explicitly model order or delay of the model. We will refer to this approach as an over-parametrization. In order to choose the number d of tapped delays in the FIR model, one may perform a nonparametric analysis of the impulse response of the data (neglecting nonlinear effects as yet). It is found that MINLIP is especially appropriate for such an over-parametrization approach and doesn't loose much efficiency as it builds in explicitly a mechanism of model complexity control and regularization, dealing with ill-posedness problems often present in such a context.

4.1 The Experimental setup

The Monotone Wiener systems from which the data is generated take the following form. The linear subsystem $H_0(q^{-1})$ are represented as fractional polynomial models as

$$H_0(q^{-1}) = \frac{B(q^{-1})}{A(q^{-1})} = \frac{b_0 + b_1q^{-1} + \cdots + b_{2m_z}q^{-2m_z}}{1 + a_1q^{-1} + \cdots + a_{2m_z}q^{-2m_p}}, \quad (34)$$

where $2m_z > 0$ and $2m_p > 0$ denote the orders of the polynomials $A(q^{-1})$ and $B(q^{-1})$. Those polynomials are chosen such that they have m_z and m_p conjugate pairs of zeros and poles respectively. The zeros of $A(q^{-1})$ are referred to as poles of $H_0(q^{-1})$, and the zeros of $B(q^{-1})$ are referred to as zeros of $H_0(q^{-1})$. In this example, we set $n_z = 2$ and $n_p = 20$. The conjugate poles and conjugate zeros are uniformly at random picked inside the unit circle (see Figure 5 for an example). In general, we see that a FIR representation of $d = 200$ is sufficient to capture the dynamics of such a system. The output nonlinearity is fixed as $f_0 : \mathbb{R} \to \mathbb{R}$ where for $x \in \mathbb{R}$ one has

$$f_0(x) = 2 + \tanh(5x+2) + 0.5\tanh(5x-3). \quad (35)$$

This function is somewhat challenging as it cannot be described as a simple saturation function, is not symmetric around any point, and has an almost zero gradient in $x = 0$. Then a monotone Winer system is constructed as

$$y_t = f_0(gH_0(q^{-1})u_t),\ \forall t = 1,\ldots,T, \quad (36)$$

where the gain $g > 0$ is chosen such that the values $\{gH_0(q^{-1})u_t\}_t$ have a unit standard deviation. The estimates of MINLIP on a time-series of length T=450, 500,550,600 - taken from the Wiener System of Fig. (5), Fig. (6) - is displayed in Fig. (6). Here a FIR approximation of $d = 200$ is used, capturing the dynamics of the system H_0 reasonably well (see Fig. (5.a)).

The following 6 different identification methods were implemented to benchmark the MINLIP against:

(1) (LS '$x-y$') In order to provide a (naive) lower-bound to the performance, a FIR identification technique based on a Least Squares (LS) argument was implemented on the signals $\{u_t\}_t$ and $\{y_t\}_t$ directly, neglecting the Wiener structure altogether.

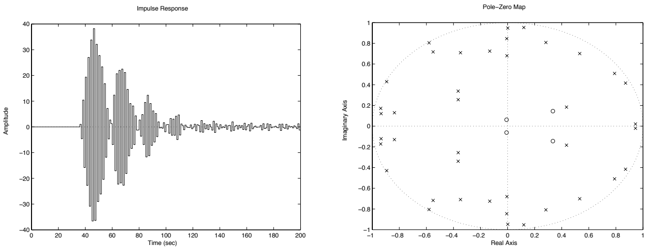

Fig. 5. *An example of a system H_0 randomly generated, with $n_p = 20$ and $n_z = 2$. Panel (a) shows the resulting impulse response (and hence a FIR approximation) up to lag $d = 200$. Panel (b) displays the conjugate poles and conjugate zeros in the complex domain using a pole-zero plot of the system. Observe the (i) considerable delay, and (ii) some poles are located close to the unit circle. This makes a inverse modeling approach unfeasible.*

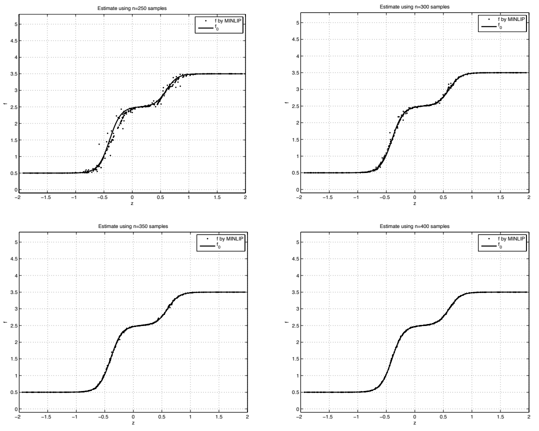

Fig. 6. *Evolution of the estimate of MINLIP when provided by signals of length (a) 450, (b) 500, (c) 550, (d) 600 samples, taken from the Wiener model described in Figure (5) and eq. (35). A FIR model of* $d = 200$ *is used to approximate the linear system, taking care of the delay as well of the (unknown) model orders.*

(2) (LS '$x - z$') In order to get an upper-bound on the performance of the identification technique, an ARX identification technique was implemented based on the (latent) intermediate signal $\{z_t = H_0(q^{-1})u_t\}_t$.

(3) (WPEM FIR) We consider a FIR model structure of sufficiently high order (here $d = 200$) such that (34) can be represented fairly well, and we let the corresponding FIR coefficients act directly as unknowns. The nonlinearity is represented as a piecewise function based on 20 fixed knots which were optimally tuned to the example at hand. Global optimization on both FIR coefficients as well as on the unknowns of the nonlinearity is performed by the Broyden-Fletcher-Goldfarb-Shannon (BFGS) method implemented in MATLAB in the `fminunc` function.

(4) (WPEM ARX) Here we implement the same approach now based on a class of ARX models representing the optimal predictors corresponding to (34). Now, we let the poles and zeros of this model class act as unknowns directly, and As before, the nonlinearity is expressed as a piecewise linear function with fixed grid points. Global optimization is performed by the BFGS method.

(5) (Greblicki2002) The smoothing approach as described in [5] is implemented as well. This method works directly on a FIR overparametrization of the (linear sub-) system, and gives reasonable estimates when sufficiently many input-samples following a stochastic (approximately white) Gaussian process are provided.

(6) (Bai2006) The last approach we benchmark against is the technique described in [19,2] using prior knowledge of the monotonicity of the output function. This technique is implemented by solving

$$\min_a \sum_{j=d+1}^{T} \left(\mathrm{sign}(y_{(j)} - y_{(j-1)}) - \widetilde{\mathrm{sign}}((\mathbf{u}_{(j)} - \mathbf{u}_{(j-1)})^T a)\right)^2, \tag{37}$$

which is in our experiments solved by the BFGS method. We found that in order to make this approach to work well one needs to resort to a smooth proxy '$\widetilde{\mathrm{sign}}$' of the discrete function '$\mathrm{sign}(z) = I(z > 0) - I(z < 0)$', making gradient information available at most unknowns in the search space. In many cases solving this problem takes substantially more resources (CPU-power, memory) compared to the other techniques.

Accuracy of an estimate is expressed in terms of the angle between the true impulse response of H_0 and the impulse response of the estimate $\hat{H}_T$. Let as classical the L_2 norm of a system H be defined as

$$\sum_{t=0}^{\infty} (H_0(q^{-1})\delta_t)^2, \tag{38}$$

where δ_τ is a time-series of all zeros except for the first location which equals one. Then the correlation of two systems H_0 and $\hat{H}_T$ may be defined as

$$d(H_0, \hat{H}_T) = \frac{\sum_{t=0}^{\infty} H_0(q^{-1})\delta_t \; \hat{H}_T(q^{-1})\delta_t}{\|H_0\|_2 \|\hat{H}_T\|_2}. \tag{39}$$

If the systems H_0 and $\hat{H}_T$ have impulse response vector h_0 and $\hat{h}$ respectively, this coefficient can be written as the Pearson correlation coefficient between those two vectors, or $\frac{h_0^T \hat{0}}{\|h_0\|_2 \|\hat{h}\|_2}$. There are 3 reasons for adopting this definition. The first is that the gain of the system H_0 is unidentifiable (hence cannot play a role in the quality measure), and the second one is that the impulse response is the common denominator for any LTI, independently of a parametrization. Thirdly, the current method concentrates on first instance only at identification of the linear system, while making predictions requires additional estimation of the static nonlinearity. As such, measuring performance based on prediction accuracy would convolute the results with performance of this reconstruction step as well.

4.2 A Noiseless Example

The first experiment is based on noiseless data. Figure (7) shows the results of the experiment, where in each iteration T samples are generated from a random system (f_0, H_0), the different identification algorithms are carried out, and their respective accuracy is computed. This iteration is performed 100 times for any $T = 300, 400, 500, \ldots, 1000$. We see from the results that the MINLIP estimator converges fast to the best achievable performance (indicated by the LS '$x-z$' approach). The WPEM algorithms give in many cases unreliable results, performing much worse on the average. Specifically, we find that it is bad practice to combine the WPEM on (overparametrized) FIR model, perhaps because global optimization often presents (numerical) problems when optimizing over so large a set of parameters. MINLIP however can handle such overparametrization quite efficiently, and does as such not require to determine the model orders and the delay of the system. This suggests the use of complexity control to give a principled tool to handle the task of model order selection.

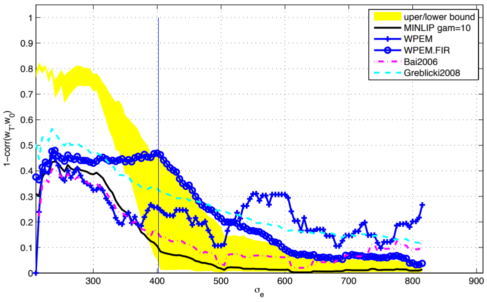

Fig. 7. *Results of the first experiment using noiseless data generated from a monotone Wiener nonlinearity as in eq. (35) and systems H_0 as in (34). Performance expressed as correlations of H_0 and $\hat{H}_T$ as in (39) are displayed for sample sizes ranging from T=210 to T=1000, using a FIR overparametrization of $d = 200$. The vertical line denotes the place where a least squares technique would exactly reconstruct H_0 if $z_t = H_0(q^{-1})u_t$ were given.*

4.3 Quantized signals

Here we investigate the use of MINLIP in case the output signal is quantized (i.e. takes a small number of different values). This task poses additional challenges as there is a direct need to handle tied output values well. We adapt the following procedure: if $y_{(i-1)} = y_{(i)}$ (tied), then we cannot compare the corresponding values $a^T\mathbf{u}_{(i-1)}$ and $a^T\mathbf{u}_{(i)}$ unambiguously. Rather, we compare $a^T\mathbf{u}_{(i-1)}$ and $a^T\mathbf{u}_{(i)}$ both with the samples $(y_j, a^T\mathbf{u}_j)$ and $(y_k, a^T\mathbf{u}_k)$ which have a strictly lower value $y_j < y_{(i-1)}, y_{(i)} < y_k$. Again by transitivity of the relation $<$ one can prune many of the relations in the final QP. In the worst case only 2 different output levels are observed on T samples, and both levels are each observed on $\frac{T}{2}$ samples. Then one has to work with $\frac{T^2}{4}$ inequality constraints rather than the $O(T)$ ones in the standard implementation. It may be argued that the theoretical account for Monotone Wiener systems as given above does not hold strictly as the steepest part ('the jumps') have an unbounded Lipschitz constant. However, as the dataset is finite, this measure is necessarily finite and the method continuous to yield good solutions. If there are no samples present 'near the jumps', the analysis hold with an appropriate function g dependent on this 'margin', and the size of the jumps.

In this example we define the output nonlinearity for any $z \in \mathbb{R}$ as

$$f_0'(z) = I(z > -0.5) + I(z > 2), \tag{40}$$

with the output taking values in the set $\{0, 1, 2\}$, and the jumps of the function occurring when $z = -0.5$ and $z = 2$. Again, the linear systems H_0 used to generate the signals are as in eq. (34), and we benchmark the MINLIP against the approaches described in the previous subsection (LS, WPEM and BAI). The results are displayed in Fig. (8).

From these results we may suggest a few guidelines. The first is that a naive LS '$x-y$' regression works surprisingly good, and it is not at all trivial to beat this one. The reason the approaches based on global optimization (i.e. WPEM, WPEM.FIR and Bai2006) do not work as good might be that the discrete nature of the identification task translates in a highly nonsmooth cost surface given to the optimizer. This experiment however suggests that MINLIP achieves a solution which is often close to the best one could hope for (indicated by the LS '$x-z$' method). The averaging approach proposed in Greblicki2008 appears fairly robust to the quantization effects as well.

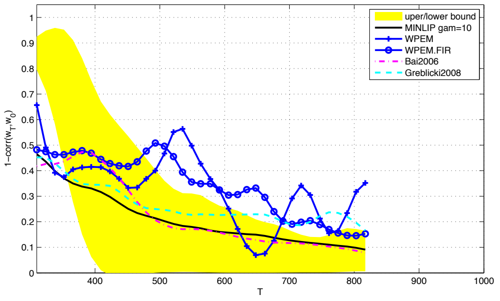

Fig. 8. *Results of the quantization experiment using noiseless data generated from a monotone Wiener nonlinearity as in eq. (40) and systems H_0 as in (34). Performance expressed as correlations of H_0 and $\hat{H}_T$ as in (39) are displayed for sample sizes ranging from $T = 210$ to $T = 1000$, using a FIR overparametrization of $d = 200$.*

4.4 The noisy case

This subsection reports results achieved with MINLIP in case the intermediate signal $\{z_t\}_t$ is perturbed by noise. The experiment is set up as before in Subsection 2.3, but the system becomes now

$$y_t = f_0\left(H_0(q^{-1})u_t + e_t\right), \; \forall t = 1, \ldots, T, \qquad (41)$$

where f_0 is as in eq. (35) and H_0 is randomly generated as in (34). The terms $\{e_t\}_t$ are zero mean white Gaussian noise with standard deviation $\sigma_e > 0$. This experiment is conceived in a slightly different manner than before. We consider a fixed number of samples $T = 500$, and let the Signal-to-Noise Ratio (SNR) vary from 0.1 to 10 (or $\sigma_e = 0.1, \ldots, 10$). As indicated in Section III, MINLIP is dependent on the choice of a suitable $\gamma > 0$, which is in turn depending on the noise variance σ_e. As this characteristic is unknown in general in practical applications, the choice of γ is to be made based on a suitable model selection technique. Actually, the problem of model selection in the context of a Wiener model is not covered as such, and prompts new questions related to information criteria, stability and consistency. For now, we use a fixed value of $\gamma = 10$ which works well in many cases. The results are displayed in graph (9). Those results indicate that the MINLIP outperforms the other techniques especially when noise is small compared to the 'informative' signal, while all techniques become arbitrarily bad when this ratio grows.

In the next experiment fix an SNR of 3 and lets see what happens to the performance of the different estimators if the given signals have an increasing length. The evolution of the average accuracy of MINLIP and the competing estimators is given in Fig. (10). From this result we see that the behavior is not too different from the noiseless case, except the fact that the WPEM and WPEM.FIR approaches are not very robust to noise, and the approach in (37) is clearly a bad choice in this case and needs additional care.

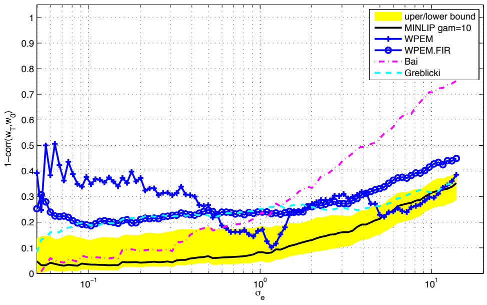

Fig. 9. *Results of the third experiment using noisy data generated from a monotone Wiener nonlinearity as in eq. (35) and systems H_0 as in (34). Performance expressed as correlations of H_0 and $\hat{H}_T$ as in (39) are displayed for a sample of sizes $T = 500$. The amount of noise varies from $\sigma_e = 0.1$ to $\sigma_e = 3$, where the linear system H_0 is rescaled such that $\sigma_z = 1$ (corresponding with SNR from 10 to 0.66). MINLIP was implemented with a fixed $\gamma = 10$, and can be seen to outperform other methods especially when the SNR is relatively high.*

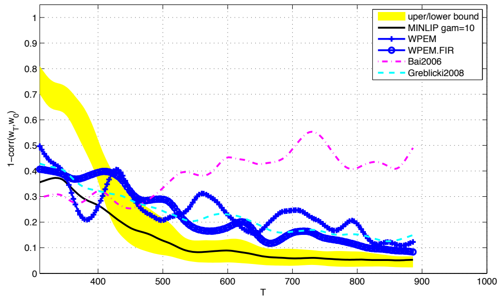

Fig. 10. *Results of the last experiment using noisy data generated from a monotone Wiener nonlinearity as in eq. (35) and systems H_0 as in (34), and SNR equal to 3. Performances of the different estimators are expressed as $1-$correlations of H_0 and $\hat{H}_T$ as in (39). The sample sizes ranges from $T = 210$ to $T = 1000$, using a FIR over-parametrization of $d = 200$.*

5 DISCUSSION

This paper studies how MINLIP works for identification of monotone Wiener systems. Theoretical as well as empirical evidence indicates the use of the estimate despite its unconventional groundings. Especially, one of the points is that this method based on model complexity control can handle FIR overparametrizations of the linear subsystem quite efficiently, implementing implicitly model order- and delay-estimation during the identification task. The crux of the method is to place model complexity control in the centre of the identification task. The hope is that this line of thinking provides novel ideas which are useful in the design and analysis of identification algorithms for more general nonlinear systems. A main open question is a theoretical study of the influence of noise in the almost consistency result.

References

[1] E.W. Bai. A blind approach to the Hammerstein-Wiener model identification* 1. *Automatica*, 38(6):967–979, 2002.

[2] E.W. Bai and J. Reyland. Towards identification of Wiener systems with the least amount of a priori information on the nonlinearity. *Automatica*, 44(4):910–919, 2008.

[3] S.A. Billings and SY Fakhouri. Identification of a class of nonlinear systems using correlation analysis. In *Institution of Electrical Engineers, Proceedings*, volume 125, pages 691–697, 1978.

[4] G.C. Goodwin and K.S. Sin. *Adaptive filtering prediction and control*. Prentice-Hall Englewood Cliffs, NJ, 1984.

[5] W. Greblicki and M. Pawlak. *Non-Parametric System Identification*. Cambridge University Press, 2008.

[6] Anna Hagenblad. *Aspects of the Identification of Wiener Models*. PhD thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, Nov 1999.

[7] L. Ljung. Analysis of recursive stochastic algorithms. *IEEE transactions on automatic control*, 22(4):551–575, 1977.

[8] L. Ljung. *System Identification, Theory for the User*. Prentice Hall, 1987.

[9] K. Pelckmans, I. Goethals, J.A.K. Suykens, and B. De Moor. On model complexity control in identification of hammerstein systems.

In *the 44th IEEE conference on Decision and Control, and the European Control Conference (CDC-EEC 2005)*, Sevilla, Spain, 2005.

[10] K. Pelckmans, T. Van Waterschoot, and J.A.K. Suykens. Efficient adaptive filtering for smooth linear fir models. In *Internal Report 10-60, ESAT-SISTA, K.U.Leuven, Belgium, submitted*. 2010.

[11] T. Söderstrom and P. Stoica. System identification, 1989.

[12] K. Tsumura and J. Maciejowski. *Optimal quantization of signals for system identification*. University of Cambridge, Department of Engineering, 2002.

[13] V. Van Belle, K. Pelckmans, J.A.K. Suykens, and Van Huffel S. Learning transformation models for ranking and survival analysis, *submitted*. 2009.

[14] J. Voros. Parameter identification of Wiener systems with discontinuous nonlinearities. *Systems and Control Letters*, 44(5):363–372, 2001.

[15] D. Westwick and M. Verhaegen. Identifying MIMO Wiener systems using subspace model identification methods. *Signal Processing*, 52(2):235–258, 1996.

[16] T. Wigren. Convergence analysis of recursive identification algorithms basedon the nonlinear Wiener model. *IEEE Transactions on Automatic Control*, 39(11):2191–2206, 1994.

[17] T. Wigren. Approximate gradients, convergence and positive realness in recursive identification of a class of non-linear systems. *International Journal of Adaptive Control and Signal Processing*, 9(4), 1995.

[18] T. Wigren. Adaptive filtering using quantized output measurements. *IEEE transactions on signal processing*, 46(12):3423–3426, 1998.

[19] Q. Zhang, A. Iouditski, and L. Ljung. Identification of Wiener system with monotonous nonlinearity. In *Proceedings of IFAC Symposium on System Identification*, page 166-171, Newcastle, Australia, 2006.