On the role of exact models in approximate modeling problems
Stoorvogel, A.A.; Weiland, S.

Published: 01/01/1996

Document Version
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 13. Sep. 2017
On the Role of Exact Models in Approximate Modeling Problems

Siep Weiland and Anton A. Stoorvogel

August, 1996

Measurement and Control Systems
Internal Report, 96 I/04


Eindhoven, August 1996
On the Role of Exact Models in Approximate Modeling Problems

Siep Weiland
Department of Electrical Engineering
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands
E-mail: s.weiland@ele.tue.nl

Anton A. Stoorvogel
Department of Mathematics
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands
E-mail: wscoas@win.tue.nl

Abstract

The behavioral theory of dynamical system is used to address a deterministic system identification problem with a newly defined measure of misfit between data and linear time-invariant systems. An approximate model identification problem is formalized using this misfit criterium. In particular, Pareto optimal models are defined as feasible trade-offs between low complexity and low misfit models. The main result of this paper provides a complete characterization of bounded misfit and bounded complexity models. It is shown that this entire class of approximate models corresponds to the set of most powerful unfalsified models of reduced data sets. The reduced data sets are derived from Hankel norm approximations of the data. The main result therefore emphasizes the relevance of the exact modeling problem for the identification of approximate systems. The set of all Pareto optimal models is characterized as a simple consequence of this result.

Keywords

System identification, Approximate modeling, Hankel operators, Behavioral theory, Linear systems.

1 Introduction

A fundamental problem in system identification is to optimally model observed data. The most common formulations of this problem involve the parametrization of a set of models whose quality is evaluated on the basis of a criterion which expresses the discrepancy between data and model. There is a well developed theory

1 Part of this research has been made possible by a grant from the European Community for the Systems Identification and Modeling Network (SIMONET).
reduced data set. In particular, this result shows the relevance of the exact modeling problem for the identification of approximate systems. It is shown that Pareto optimal models are obtained as a simple corollary from this result (see Theorem 5.3 below).

The paper is organized as follows. Some notation is introduced in the next section. A formal problem statement is presented in Section 3. Section 4 treats representations of $\ell_2$ systems. The main results are presented in Section 5 and further discussed in Section 6. In the latter section we also consider a number of generalizations of the misfit function treated in this paper. Conclusions are given in Section 7. Because of space limitations we do not include proofs of the main results in this paper. We refer to [12] for the proofs and a more extensive discussion of the results of this paper.

2 Notation

For $T \subseteq \mathbb{Z}$ we denote by $\ell_2(T, \mathbb{R}^q)$ the set of functions $w : T \to \mathbb{R}^q$ for which

$$
\|w\|_2 := \sum_{t \in T} \|w(t)\|^2
$$

is finite. We write $\ell_2^+$ for $\ell_2(Z_+, \mathbb{R}^q)$ and $\ell_2^-$ for $\ell_2(Z_-, \mathbb{R}^q)$. $\mathcal{L}_2$ denotes the space of all complex valued functions which are square integrable on the unit disc $\mathbb{D} := \{z \in \mathbb{C} \mid |z| = 1\}$. The Hardy space $H^2_\infty$ consists of all square integrable functions on $\mathbb{D}$ with analytic continuation outside the unit circle (including $\infty$). The orthogonal complement of $H^2_\infty$ in $\mathcal{L}_2$ is denoted $\mathcal{H}^2_\infty$. $\mathcal{H}^2_\infty$ and $\mathcal{H}^\infty_\infty$ denote the Hardy spaces of complex valued functions which are bounded on the unit circle with analytic continuation in $|z| > 1$ and $|z| < 1$, respectively. The prefix $\mathcal{F}$ is used to denote rational elements of Hardy spaces. The canonical projections from $\mathcal{L}_2$ to $H^2_\infty$ and $\mathcal{H}^2_\infty$ will be denoted by $\Pi_+$ and $\Pi_-$, respectively.

3 The identification problem

Consider a finite set of real valued multivariable time-series

$$w_i : Z_+ \to \mathbb{R}^q, \quad i = 1, \ldots, N$$

(3.1)

where $N > 0$ is the number of observed time-series and $q > 0$ is the dimension of the signal space. It will be assumed that $w_i \in \ell_2^+$ for $i = 1, \ldots, N$. Using the Hilbert space isomorphism between $H^2_\infty$ and $\mathcal{H}^2_\infty$ it follows that the data (3.1) can compactly be represented by the $\mathcal{H}^2_\infty$ function

$$W(z) := [\hat{w}_1(z) \quad \hat{w}_2(z) \quad \cdots \quad \hat{w}_N(z)]$$

where $\hat{w}_i$ denotes the $z$-transform of $w_i$, i.e.,

$$\hat{w}_i(z) := \sum_{t=0}^{\infty} w_i(t)z^{-t}.$$
**Definition 3.3 (Misfit)** The misfit between a model $B \in B_2$ and the data $W$ is defined as

$$d(B, W) := \sup \left\{ \frac{\langle W \lambda, v \rangle}{\|\lambda\|_2 \|v\|_2} \mid v \in \mathbb{H}^+ \right\}$$

(3.3)

where the norms are the standard norms in the Hilbert spaces $\mathbb{H}^2_\tau$ and $\mathbb{H}_2$ and $B^\perp$ is the orthogonal complement of $B$ in $\mathbb{H}^2_\tau$.

This misfit function has the following interpretation. The set

$$B_0 := \overline{\text{span}}\{s^i w_i \mid i = 1, \ldots, N; \; t \in \mathbb{Z}_+\}$$

(3.4)

defines a linear, time invariant $\ell_2$-system which is unfalsified by the data (3.1) in the sense that the data $w_i \in B_0$ for $i = 1, \ldots, N$. By associating with $W$ a multiplicative operator $W : \mathbb{H}^2_\tau \to \mathbb{L}_2$, it is easily seen that $B_0$ is the closure of $\Pi_+ W \mathbb{H}^2_\tau$ and that the misfit $d(B_0, W) = 0$. For any $B \in B_2$, the values of $(w, v)$ with normalized elements $w \in B_0$ and $v \in B^\perp$ therefore indicate to what extent linear functionals implied by $B$ fail to explain the unfalsified model $B_0$ of the data. The misfit (3.3) can therefore equivalently be viewed as a distance measure between the exact model $B_0$ and the $\ell_2$-system $B \in B_2$. To make this precise, consider $B_0$ and equip $B_0$ with the data weighted norm

$$\|w\|_W := \inf \left\{ \|x\| \mid x \in \mathbb{H}^2_\tau, \; w = \Pi_+ W x \right\}. $$

Then it is easy to show that

$$d(B, W) = \sup \left\{ \frac{\langle w, v \rangle}{\|w\|_W \|v\|_2} \mid w \in B_0, \quad v \in B^\perp \right\}. $$

(3.5)

The misfit therefore incorporates a weighting along the principal axes of the image of $\Pi_+ W$ under $\mathbb{H}^2_\tau$ and can be interpreted as a data-weighted distance measure between $B_0$ and $B$.

**Remark 3.4** It is interesting to point out the relation between the expression (3.5) and the gap between the two closed subspaces $B \subseteq \ell_2^+$ and $B_0 \subseteq \ell_2^+$. The gap $[2, 3]$ is defined as

$$g(B, B_0) := \sup \left\{ \frac{\langle w, v \rangle}{\|w\|_2 \|v\|_2} \mid w \in B_0, \quad v \in B^\perp \right\}. $$

(3.6)

Compared with (3.5), the gap has therefore a different normalization of elements in $B_0$.

**Remark 3.5** We would like to emphasize the importance of this data-weighting for applications of the misfit (3.3) in system identification. Indeed, if we consider the example of an observed scalar valued time series $w(t) = \lambda_1^0 + \epsilon \lambda_2$, with $t \in \mathbb{Z}_+$, $|\lambda_1| < 1, \; i = 1, 2$ and $\epsilon > 0$, then $w \in \ell_2^+$ and the linear span of the time series $\lambda_1^0$ and $\lambda_2$ defines the exact model $B_0$ in (3.4). It has complexity $c(B_0) = (0, 2)$. A lower order approximate model obtained by minimizing the gap (3.6) over all $B \in B_2$ with complexity $c(B) = (0, 1)$ would therefore result in a model which does not depend on the value of $\epsilon$. That is, for small $\epsilon > 0$, the relative weight of the component $\lambda_1^0$ of the data sequence $w$ is not discriminated by the gap criterion. However, in the misfit-criterion (3.3) it is.

**Remark 3.6** Clearly, $d(B, W) \geq 0$ and $d(B, W) = 0$ if and only if $w_i \in B$ for $i = 1, \ldots, N$. Further, it is immediate from Definition 3.3 that $B_1 \subseteq B_2$ implies that $d(B_1, W) \geq d(B_2, W)$.

**Definition 3.7 (Pareto optimality)** A model $B \in B_2$ is called Pareto-optimal for the data $W$ if for all $B' \in B_2$ the implications

$$c(B') \leq c(B) \Rightarrow d(B', W) \geq d(B, W)$$

$$d(B', W) \leq d(B, W) \Rightarrow c(B') \geq c(B)$$

hold simultaneously.

Hence, a model is Pareto optimal with respect to the data (3.1) if all models of smaller complexity have larger misfit, and all models of smaller misfit have larger complexity.

**4 Kernel representations**

Before presenting the main results we address in this section the problem of representing elements of the model class $B_2$. It is shown in [10, 11] that $B_2$ consists of those systems which can be represented as the kernel of a rational operator. We show here that the complexity and the misfit function introduced in the previous section admit simple characterizations in terms of kernel representations.

Let $\Theta \in \mathcal{R} \mathcal{H}_\infty$. Then $\Theta$ defines a map $\Theta : \mathcal{H}^+_\infty \to \mathbb{L}_2$ defined by the multiplication $[\Theta w](z) = \Theta(z)w(z)$. Associate with each such $\Theta$ a model

$$\hat{B}_{\Theta}(\Theta) := \{w \in \mathcal{H}^+_\infty \mid \Theta w \in \mathcal{H}^+_\infty \} = \ker \Pi_+. $$

Let $B_{\Theta}(\Theta)$ denote its inverse $z$-transform. It is then easy to see that $B_{\Theta}(\Theta)$ belongs to $B_2$. In fact, the following result claims that every model in $B_2$ admits such a representation.

**Theorem 4.1 (Kernel representation theorem)** $B \in B_2$ if and only if there exists $\Theta \in \mathcal{R} \mathcal{H}_\infty$ such
that $B = B_{\text{exc}}(\Theta)$. Moreover, every $B \in \mathcal{B}_2$ can be represented as $B = B_{\text{exc}}(\Theta)$ with $\Theta \in \mathcal{RH}_\infty$ co-inner, i.e. $\Theta\Theta^* = I$ where $\Theta^*(z) := \Theta^T(z^{-1})$.


The Hardy space $\mathcal{RH}_\infty$ therefore provides a parametrization of the model class $\mathcal{B}_2$. The next results completely characterize the complexity and misfit in terms of kernel representations.

**Theorem 4.2** Let $\Theta \in \mathcal{RH}_\infty$ be a co-inner kernel representation of $B \in \mathcal{B}_2$. Then the following statements are equivalent

1. $B$ has complexity $c(B) = (m,n)$.
2. $m = q - \text{rank}(\Theta)$ and $n = \text{deg}(\Theta)$, where $\text{deg}(\Theta)$ denotes the McMillan degree of $\Theta$.

**Proof.** See [12].

To characterize the misfit function, define the Hankel norm of $G \in \mathcal{L}_\infty$ as

$$\|G\|_H := \sup_{x \in \mathcal{H}_2^+} \|\Pi_x G x\|_2$$

Thus $\|G\|_H$ is the induced norm of the Hankel operator $\Gamma_G := \Pi_+ G \Pi_- \text{ acting on } \mathcal{H}_2^+$ as a multiplicative operator. Using this definition, the misfit function defined in 3.3 is characterized as follows.

**Theorem 4.3** Let $\Theta \in \mathcal{RH}_\infty$ be a co-inner kernel representation of $B \in \mathcal{B}_2$. Then

$$d(B, W) = \|\Theta W\|_H$$

**Proof.** See [12].

Hence, the misfit is equal to the Hankel norm of the $\mathcal{L}_\infty$ function $\Theta W$.

**5 Main results**

The important notion of a most powerful unfalsified model was introduced by Willems (see e.g. [9]) and plays a crucial role in the sequel to characterize both exact and Pareto optimal approximate models for a given data set. An element $B \in \mathcal{B}_2$ is said to be an unfalsified model for $W \in \mathcal{H}_2^+$ if $d(B, W) = 0$. It is called the most powerful unfalsified model for $W$ if in addition $B' \in \mathcal{B}_2$ with $d(B', W) = 0$ implies that $B \subseteq B'$. The most powerful unfalsified model is unique whenever it exists and can be interpreted as the smallest $\ell_2$ system that explains the data, i.e., the smallest covering of $W$ by elements in $\mathcal{B}_2$. If it exists, we will write

$$B = \text{mpum}(W)$$

to denote the most powerful unfalsified model corresponding to $W$.

If the data $W$ is rational then the most powerful unfalsified model actually exists and is easily obtained:

**Theorem 5.1** (Exact models) If $W \in \mathcal{RH}_\infty^+$ then there exists a left coprime factorization $W = \Theta_0^{-1} \Psi_0$ over $\mathcal{RH}_\infty$. For any such factorization there holds that

1. $B_{\text{exc}}(\Theta_0)$ is the most powerful unfalsified model for $W$.
2. $B_0 = B_{\text{exc}}(\Theta_0)$.
3. $B_{\text{exc}}(\Theta)$ is an unfalsified model for $W$ if and only if $\Theta = \Lambda \Theta_0$ with $\Lambda \in \mathcal{RH}_\infty$.

**Proof.** See [12] or [1].

Hence, all left-coprime factorizations of $W$ define kernel representations of the most powerful unfalsified model. Moreover, all unfalsified models of $W$ are parametrized by left multiplication of a left-coprime factor of $W$ with rational operators in $\mathcal{RH}_\infty$. The second main result of this paper characterizes the set of all autonomous systems of given complexity and prescribed misfit level.

**Theorem 5.2** (Approximate models) Let $B \in \mathcal{B}_2$ and $\sigma > 0$ be given and assume that $W \in \mathcal{RH}_\infty^+$. Let $\sigma_1 \geq \ldots \geq \sigma_n > 0$ denote the ordered Hankel singular values of $W$. Then the following statements are equivalent

1. $B$ has complexity $c(B) = (0,k)$ and misfit $d(B, W) \leq \sigma$.
2. $\sigma \geq \sigma_{k+1}$ and there exists $W_k \in \mathcal{RH}_\infty^+$ such that $\|W - W_k\|_H \leq \sigma$, $\deg(W_k) \leq k$ and $B = \text{mpum}(W_k)$.

**Proof.** See [12].

Theorem 5.2 characterizes an entire set of feasible complexity and misfit combinations of dynamical systems for a given data set $W$. In words, it claims that the class of approximate autonomous models of the data set $W$ is characterized as the class of exact models of data sets $W_k$ which are obtained as Hankel norm approximants of $W$. We believe that this result has conceptually interesting consequences. Instead of finding
approximate models directly from data, this result suggests to determine a related data set, $W_k$, whose most powerful unfalsified model has guaranteed properties of complexity and misfit.

Pareto optimal models are easily derived using the characterization of Theorem 5.2. The following result shows that Pareto optimal models precisely correspond to the most powerful unfalsified models of optimal Hankel norm approximants of the data $W$.

**Theorem 5.3 (Pareto optimal models)** Let $W \in \mathcal{RH}^+_\infty$ be given and let $B \in \mathcal{B}_2$ have complexity $c(B) = (0, k)$ and misfit $d(B, W) = \sigma$. Assume moreover that the Hankel singular value $\sigma_k(W) > \sigma_{k+1}(W)$. Then the following statements are equivalent:

1. $B$ is Pareto optimal for the data $W$,
2. $\sigma = \sigma_{k+1}$,
3. there exists an optimal Hankel norm approximant $W_k$ of $W$ of McMillan degree $k$ such that $B = \text{mpum}(W_k)$.

**Proof.** See [12].

### 6 Discussion and generalizations

**Remark 6.1** The assumption on rationality of $W$ may seem very restrictive at first sight. However, we remark that this assumption is satisfied for important signal sets like polynomial-exponential data, impulse responses of linear time-invariant lumped systems, finite support time-series or any finite set of frequency response samples. Also, the rationality of the Laplace transformed data is a common assumption in many problems in realization theory and the theory of rational interpolation.

**Remark 6.2** In the presented setting no assumptions were made on noise characteristics. We emphasize that finite data sequences and finite number of frequency domain measurements always imply the rationality of $W$ irrespective of noise models. Specific assumptions on noise are therefore not necessary.

**Remark 6.3** A conceptual and constructive procedure for the computation of bounded complexity and bounded misfit models can be derived from the results of Section 5 and the results of Glover in [4] as follows. Given $W \in \mathcal{RH}^+_\infty$, $k > 0$ and $\sigma > 0$ one can compute the (ordered) Hankel singular values $\sigma_i$ of $W$ and verify whether $\sigma \geq \sigma_{k+1}$. If $\sigma_k > \sigma > \sigma_{k+1}$ define the augmented operator

$$W_\sigma(z) := \begin{pmatrix} W(z) & 0 \\ 0 & \sigma(z+1)/z \end{pmatrix},$$

otherwise take $W_\sigma = W$. Let $W_k^\sigma$ be an optimal Hankel norm approximant of McMillan degree $k$ to $W_\sigma$ (see [4]) and let $W_k$ be the truncation of $W_k^\sigma$ to the Birkhoff operator. $W_k$ will then satisfy $\|W - W_k\|_B \leq \sigma$. Next, let $W_k = \Theta_k^{-1} \Psi_k$ be a left-coprime factorization of $W_k$ over $\mathcal{RH}^\sigma_\infty$ and put $B = B_{\text{mpum}}(\Theta_k)$. Then $B$ belongs to $\mathcal{B}_2$ and satisfies the criteria stated in Theorem 5.2.1. Moreover, if $\sigma$ is strictly larger than $\sigma_{k+1}$ then $B$ is Pareto optimal if and only if $\sigma = \sigma_{k+1}$.

**Remark 6.4** In Theorem 5.2 and Theorem 5.3 only autonomous models in $\mathcal{B}_2$ are considered. Similar results for non-autonomous systems have not been derived yet. However, since the definition of misfit warrants that $d(B', W) \leq d(B, W)$ whenever $B \subseteq B'$, we have that (in the notation of Theorem 5.2) any unfalsified model $B'$ of $W_k$ satisfies

$$d(B', W) \leq \sigma.$$

Hence, any such model has a guaranteed upperbound on the misfit. In particular, this class of unfalsified models includes non-autonomous systems. Precisely, if $W_k = \Theta_k^{-1} \Psi_k$ is a left-coprime factorization over $\mathcal{RH}^\sigma_\infty$ of $W_k$, then it follows from Theorem 5.1 and Theorem 4.2 that for any $\Lambda \in \mathcal{H}_\infty^-$ the system

$$B' = B_{\text{mpum}}(\Lambda \Theta_k)$$

admits $m = q - \text{rank}(\Lambda \Theta_k)$ inputs while it has a guaranteed misfit level $d(B', W) \leq \sigma$. Note that such a system may have a kernel representation with McMillan degree strictly smaller than $k$ by choosing $\Lambda$ appropriately.

**Remark 6.5** The misfit function (3.3) is defined in terms of the standard inner product on $\mathcal{L}_2$ and involves normalization of $\lambda$ and $v$ in terms of their standard norms. As an alternative, consider the misfit criterion

$$d_{\infty}(B, W) := \sup \left\{ \frac{(W\lambda, v)}{\|\lambda\|_\infty \|v\|_2} \mid v \in B^\perp, \Lambda \in \mathcal{H}_\infty^- \right\}. \quad (6.1)$$

Here, $\| \cdot \|_\infty$ denotes the $\mathcal{H}_\infty^-$ norm. In fact, if the data consists of one observed time series ($N = 1$), then the misfit criterion (6.1) is equal to the distance between $W$ and $B$ measured in the $\mathcal{H}_\infty^+$-norm. Precisely, if $N = 1$ then

$$d_{\infty}(B, W) = \inf_{W' \in \mathcal{B}} \|W - W'\|_2.$$

The latter criterion has been investigated in [5, 6] for the purpose of system identification.

**Remark 6.6** As another generalization, we can replace the $\ell_2$ setting of the identification problem treated in this paper by an $\ell_\infty$ framework. Let $\ell_\infty$ denote the real normed linear space of all vector valued, magnitude bounded sequences $w : \mathbb{Z}_+ \to \mathbb{R}$. 
Let \( c_0 \) denote the subset of \( \ell_{\infty} \) consisting of all sequences \( w \in \ell_{\infty} \) which vanish in the limit, i.e. for which \( \lim_{t \to -\infty} w(t) = 0 \). Assume that the data \( (3.1) \) belongs to \( c_0 \) and consider the model class \( \mathbb{B}_{\infty} \) consisting of all linear time-invariant and complete subsets of \( c_0 \). The 'orthogonal complement' \( B^\perp \) of \( B \in \mathbb{B}_{\infty} \) is defined as the set of all bounded continuous functionals which vanish at \( B \). Equipped with its induced norm, \( B^\perp \) becomes a subspace of the normed linear space \( \ell_1 \) and we define the misfit function

\[
d_1(B, W) := \sup \left\{ \frac{\|W \ast \lambda, v\|}{\|\lambda\|_1 \cdot \|v\|_1} \right\} = \sup \left\{ \frac{\|W \ast \lambda, v\|}{\|\lambda\|_1} \right\} \quad (6.2)
\]

where \( W := [w_1, \ldots, w_N] \) is defined in the time-domain. We claim that the misfit (6.2) is again characterized as the induced norm of a Hankel operator. That is, if \( \Theta \in \ell_1^* \) has norm \( \|\Theta\|_1 = 1 \), then \( \Theta \) defines a normalized kernel representation of a system \( B \in \mathbb{B}_{\infty} \) by putting

\[
B = B_{\text{ker}}(\Theta) := \{w \in c_0 \mid [\Theta \ast w](t) = 0 \text{ for all } t \in \mathbb{Z}_+\}.
\]

In this setting, the misfit is characterized as follows.

**Theorem 6.7** For all \( W \in c_0 \) and \( B \in \mathbb{B}_{\infty} \)

\[
d_1(B, W) = \|\Theta \ast W\|_{H_1}
\]

where \( \|\cdot\|_{H_1} \) denotes the \( \ell_1 \) induced norm of the Hankel operator, i.e. if we associate with \( G \in \ell_1(\mathbb{Z}, \mathbb{R}^{x4}) \) the convolution operator \( G : \ell_1^* \to \ell_1^* \) defined as

\[
G(u) := G \ast u
\]

then

\[
\|G\|_{H_1} := \sup_{u \in \ell_1^*} \frac{\|G(u)\|_1}{\|u\|_1}.
\]

**7 Conclusions**

For a model class of \( \ell_2 \) dynamical systems we proposed definitions for misfit and system complexity which were shown to have simple characterizations in terms of rational kernel representations. These characterizations have been used to parameterize a class of optimal approximate models as exact models of related data sets. Pareto optimal models were defined and it has been shown how a set of Pareto optimal models can be obtained by determining optimal Hankel norm approximants derived from the data. In particular, these Pareto optimal models are characterized as the most powerful unfalsified models of Hankel reduced data sets. A conceptual procedure for the computation of these models has been given and various generalizations of the misfit criterion were discussed.

**References**


