Joint 48th IEEE Conference on Decision and Control and
28th Chinese Control Conference
Shanghai, P.R. China, December 16-18, 2009

Parametrization Invariant Covariance Quantification in Identification of Transfer Functions for Linear Systems[†]

Tzvetan Ivanov[*] and Michel Gevers[*]

Abstract— **This paper adresses the variance quantification problem for system identification based on the prediction error framework. The role of input and model class selection for the auto-covariance of the estimated transfer function is explained without reference to any particular parametrization. This is achieved by lifting the concept of covariance from the parameter space to the system manifold where it is represented by a positive kernel instead of a positive definite matrix. The Fisher information metric as defined in information geometry allows an interpretation as a signal-to-noise ratio weighted standard metric after embedding the system manifold in the Hardy space of square integrable analytic functions. The reproducing kernel of the tangent space with respect to this metric is shown to provide an asymptotically tight lower bound for the positive kernel representing the covariance at the system which generated the input-output data.**

Keywords: System Identification, Auto Covariance Quantification, Information Geometry, Fisher Information metric, H^2 space, Real Rational Module, Christoffel-Darboux, Reproducing Kernel

I. Introduction

A typical problem considered in system identification (SYSID) is the design of estimators trying to recover a discrete time linear time invariant (LTI) system based on a noise corrupted output sequence resulting from a known input sequence. In a parametric framework one assumes a priori knowledge about the system which restricts the uncertainty about the unknown system to a set that can be effectively parameterized. Examples include submanifolds of the space of stable causal transfer functions having constant McMillan degree with additional constraints such as model structures of Box-Jenkins (BJ), Autoregressive Exogeneous (ARX), Output-Error (OE), or Finite Impulse Response (FIR) type. Given a statistical model, the Fisher-Information matrix arising from the Fisher Information metric (FIM) provides a theoretical lower bound for the covariance of any unbiased estimator of the parameter vector. However, in general, given finitely many input-output samples there is no guarantee that an unbiased estimator exists nor that this bound, also known as the Cramér-Rao lower bound (CRLB), is tight, i.e., can be achieved. Moreover the notion of unbiasedness and performance measures such as covariance matrices is not preserved under coordinate transformations [1]. In this paper we contribute a new and natural notion of covariance which is based solely on the transfer function. Formally the covariance of a transfer function estimator is defined as a two-variable function which maps any two frequencies to the correlation of the deviation from the mean at those frequencies; this is precisely the auto-covariance if one regards the transfer function estimator as a stochastic process whose sample paths are functions on the unit circle. The covariance given by this two-variable function is Hermitian and positive in the sense of E.H. Moore [2] and hence defines a positive kernel reproducing a space of functions on the unit circle. Similar to the cone of all covariance matrices, the set of all positive kernels admits a partial ordering relation. With respect to this partial ordering we contribute a coordinate free version of the CRLB which states that the covariance of any unbiased transfer function estimator is bounded from below by the reproducing kernel with respect to the FIM on the tangent space of the system manifold at the point given by the transfer function from which the data samples originated. The sensitivity space of the prediction error previously used in [3] and [4] can thus be bypassed and replaced by the tangent space, which is a well studied object [5], [6]. Another benefit of this is that our approach is unencumbered by the sensitivity space and thus the resulting variance error quantification is not restricted to Prediction Error Methods (PEM). The analysis of estimators is carried out in the simplifying scenario where the number of data samples is large in comparison to the number of parameters to be estimated and thus can be approximated by the asymptotic behavior of the FIM, which captures the average information per data sample and thus mimics the independent identical distributed (i.i.d.) case, which is well studied in statistical literature.

This paper is structured as follows: In Section II we informally introduce the notion of covariance for a transfer function and discuss how it is influenced by the dynamics of the given true system, the chosen model class and the chosen input sequence. In Section III we take a step back and discuss the general problem of statistical inference on function spaces without any restriction and without any reference to a particular parametrization of those functions. We introduce the FIM in a coordinate free manner and prove an abstract version of the CRLB providing a lower bound for the positive kernel which represents the covariance instead of traditional positive definite matrices used in classical parameter estimation. In Section IV we demonstrate how the abstract CRLB applies in the context of SYSID and point out promising links to the theory of real rational modules

[†]This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

[*]Tzvetan Ivanov and Michel Gevers are with the Center for Systems Engineering and Applied Mechanics (CESAME) Universite Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. (http://www.inma.ucl.ac.be/~{ivanov,gevers}

and Hardy spaces. In Section V we draw conclusions and indicate future work.

Notation: $(\cdot)^*$ adjoint of complex matrix. $\nabla(\cdot)$ gradient of a function. $E_P[\cdot]$ expectation operator. $\mathbb{T}$ unit circle. $H^2(\mathbb{T})$ Hardy space. $\mathbb{K}$ field of real or complex numbers.

II. Motivational Examples

In this section we stress by means of examples that, when the data have been collected in open loop, the covariance of a transfer function estimator depends on three things: the selected model class, the system which generated the data, and the input which was chosen to identify the system. We choose to postpone the technical assumptions to Section IV to keep the discussion here informal.

Assume we are given samples $x^N = (x_1, x_2, \ldots, x_N)$ with $x_t = (u_t, y_t)$ of a process satisfying

$$y_t + 0.25\, y_{t-2} = 0.25\, u_{t-2} + v_t + 0.25\, v_{t-2}, \quad (1)$$

where v_t is i.i.d. with zero mean and unit variance. This can be written in OE form as

$$y = Gu + v \quad \text{with} \quad G = \frac{1}{1+4z^2}, \quad (2)$$

with G having poles at $z = \pm\frac{\jmath}{2}$. The parametrization of the model class $\mathcal{G}$ is[1]

$$\Pi : \Theta \to \mathcal{G}, \theta \mapsto \frac{\theta_1 z^{-1} + \theta_2 z^{-2}}{1 + \theta_3 z^{-1} + \theta_4 z^{-2}}, \quad (3)$$

and the PEM estimator of G given x^N is denoted by $\hat{G}_N$. In the following we assume $\Phi_u = 1$ that is the input has a unit spectrum. We define the covariance of $\hat{G}_N$ evaluated at the complex frequencies $z, w \in \mathbb{T}$ by[2]

$$\mathrm{Cov}(\hat{G}_N)(z,w) = E[(\hat{G}_N - G)(z)(\hat{G}_N - G)^*(w)]. \quad (4)$$

In Section IV we will show how to explicitly calculate an approximation $\mathrm{Cov}(\hat{G}_N)(z,w) \approx N^{-1}K(z,w)$ of the covariance for large N which, e.g., in this case yields

$$K(z,w) = \frac{15w(w+z)}{(4+w^2)(1+4z^2)}. \quad (5)$$

The function $K(z,z)$ represents the variance at the complex frequency $z \in \mathbb{T}$ and is easily verified to be positive and symmetric such that $K(z,z) = K(z^{-1}, z^{-1})$. We will derive an abstract result in Section III, Theorem 4, where we will explain how $\mathrm{Cov}(\hat{G}_N)(z,w)$ depends on both the true system G as well as the model class $\mathcal{G}$.

If, for example, we observe data $\tilde{x}^N$, generated by an OE model of the same form $\tilde{y} = \tilde{G}u + v$ with

$$\tilde{G} = \frac{1}{\frac{1}{2} + z + z^2} \quad \text{having poles at} \quad -\frac{1}{2} \pm \frac{\jmath}{2}, \quad (6)$$

we obtain a different covariance which is approximated by $N^{-1}\tilde{K}(z,w)$ for large N where again $\tilde{K}$ can be calculated by the methods given in Section IV; Fig.1 compares $K(z,z)$ and $\tilde{K}(z,z)$.

[1]Here $\Theta \subseteq \mathbb{R}^4$ is open and excludes pole-zero cancellations.

[2]Note in particular that $\mathrm{Cov}(T)(z,z)$ gives $E[|(\hat{G}_N - G)(z)|^2]$, i.e., the variance at z.

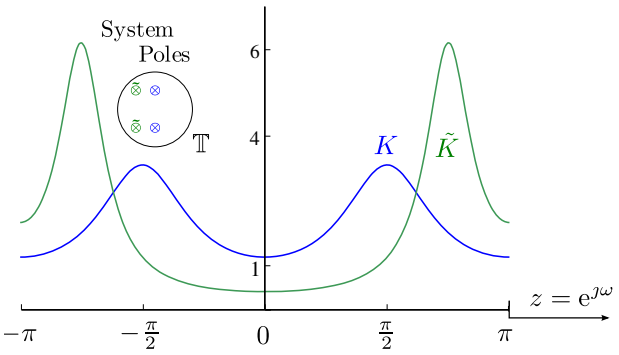

Fig. 1. $K(z,z)$ and $\tilde{K}(z,z)$ corresponding to G and $\tilde{G}$. Note that $\tilde{K}$ has higher peaks than K since the system poles, denoted by $\tilde{\otimes}$, of $\tilde{G}$ are closer to the unit circle $\mathbb{T}$ than the poles of G which are denoted by $\otimes$.

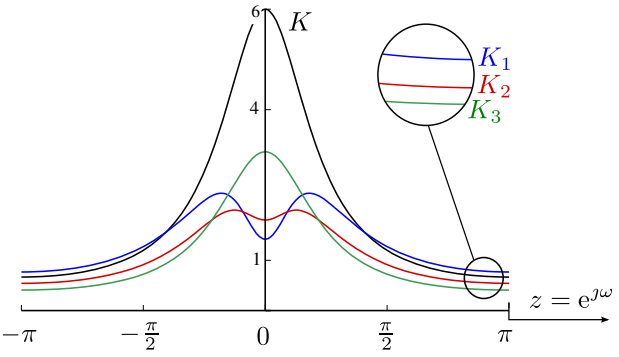

Fig. 2. K corresponds to an input with $\Phi_u = 1$ whereas K_i correspond to a periodic input sequence $u = (u_k)$ with $u_k = \cos(\omega_i k)$ with $\omega_1 = 0.3$, $\omega_2 = 0.4$ and $\omega_3 = 0.6$. Note that in a high frequency band $K_3 \le K_2 \le K_1$ whereas in the low frequency band $K_1 \le K_2 \le K_3$.

In order to facilitate our discussion regarding the influence of the input spectrum Φ_u on $\mathrm{Cov}(\hat{G}_N)$ we turn to a simpler two parameter case with

$$G = \frac{1}{z - 1/2} \quad \text{and} \quad \Pi : \theta \mapsto \frac{\theta_1}{z - \theta_2}. \quad (7)$$

Let $q = (z - 1/2)^2$ and define the Chebyshev moments of $\Phi_u \cdot |q|^{-2}$, i.e.,

$$\mu_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\jmath k\omega} \frac{\Phi_u(\omega)}{|q(\mathrm{e}^{\jmath\omega})|^2}\, \mathrm{d}\omega \quad k = 0, 1, 2, \ldots. \quad (8)$$

Then there holds

$$E[|\hat{G}_N - G|^2(\mathrm{e}^{\jmath\omega})] \approx \tfrac{2}{N}\tfrac{\mu_0 - \mu_1\cos(\omega)}{(\mu_0^2 - \mu_1^2)\cdot|\mathrm{e}^{\jmath\omega} - 1/2|^4}. \quad (9)$$

Thus the variance depends on q, which is the squared denominator of G, and on the input spectrum Φ_u only through the first two moments μ_0, μ_1 of $\Phi_u/|q|^2$. In equation (8) we allow point spectral measures resulting from periodic excitations where Φ_u is to be intepreted in the distributional sense. The accuracy of the estimator in a specific frequency band increases, i.e., the variance decreases, with the amount of input power in that band; see Fig.2.

III. Geometric Properties of Statistical Spaces

We separate this section from the SYSID context because in what follows it will not matter whether we seek to estimate a transfer function of a LTI system or a general function. We restrict our attention to the case where the function to be estimated is a priori known to reside in a finite dimensional smooth manifold. In particular we assume that the set of functions in which the function estimator takes its values can be parametrized by a finite number of real numbers. In Section III-A we define the notion of regular statistical spaces for which we define the FIM in Section III-B. The abstract version of the CRLB for function estimation is given in Section III-C and its asymptotic properties are discussed in Section III-D.

A. Regular Statistical Spaces

In terms of mathematical structure a statistical space is a triple $(X, \mathcal{X}, \mathcal{P})$ where $\mathcal{P}$, the so called *population*, is a family of probability measures $\mathcal{P} = \{P_\theta\}_{\theta\in\Theta}$ on a common measurable space $(X, \mathcal{X})$ where X is called *sample space*. The set Θ is called *parameter space* and often carries additional structure such as being a differentiable manifold or an open subset of $\mathbb{R}^n$. Any random variable $\xi : \Omega \to X$ defined on some probability space $(\Omega, \mathcal{A}, P)$ is called *sample* if its law obeys $P\xi^{-1} \in \mathcal{P}$. Loosely speaking, given an *observation* x, i.e., a realization of a sample ξ, the statistician tries to make inferences about θ, such as estimating the parameter by a function $\hat{\theta}(x)$. More generally one tries to make inference about *functionals* $f : \Theta \to \mathcal{S}$ using *estimators* $T : X \to \mathcal{S}$ where $\mathcal{S}$ can be any set.

In the following we will impose some regularity conditions on the statistical space $(X, \mathcal{X}, \mathcal{P})$ (see e.g. [1, p. 1567] for details). We assume Θ is an open subset of $\mathbb{R}^n$ and the parametrization $\pi : \Theta \to \mathcal{P}$, mapping a parameter $\theta \mapsto P_\theta$ to the corresponding probability measure, is bijective. This turns $\mathcal{P}$ into an n-dimensional manifold. We assume $\mathcal{P}$ is dominated by a σ-finite measure ν on $\mathcal{F}$. This means $\mathcal{P}$ can be identified with a family of ν-densitities $\{p(\cdot, \theta)\}_{\theta\in\Theta}$. In addition to $\mathrm{d}P_\theta = p(\cdot, \theta)\mathrm{d}\nu$ we assume that the standard regularity conditions are met.[3] The regularity conditions ensure that Definition 1 introducing below the Fisher *Information metric* on $\mathcal{P}$ makes sense. Moreover they are necessary for Theorem 2, which justifies the term information for this metric since it allows bounding the variance of all unbiased estimators from below.

B. Abstract Fisher Information Metric

The Riemannian structure on the probability manifold has been extensively studied by Amari et al [7].

[3]Those conditions are:

1) The following functions on X with $i = 1, \ldots, n$ are well defined
$$X \to \mathbb{R}, x \mapsto \partial \log p(x, \theta)/\partial\theta_i$$
$\mathbb{R}$-linearly independent and $\mathrm{d}P_\theta$-integrable.
2) $p(x, \theta)$ is a smooth function of θ for all $x \in X$ such that partial derivatives $\frac{\partial}{\partial\theta_i}$ and integration with respect to $\mathrm{d}\nu$ of $p(x, \theta)$ can always be interchanged.

Definition 1 For any $x \in X$ the *log-likelihood* function $\ell_x : P_\theta \mapsto \log p(x, \theta)$ defines a smooth map $\mathcal{P} \to \mathbb{R}$. We denote the tangent space of $\mathcal{P}$ at a point $P \in \mathcal{P}$ by $T_P\mathcal{P}$. The defining property of the Fisher Information metric (FIM) $\mathfrak{g}$ is given by

$$\mathfrak{g}_P(u, v) = E_P[\mathrm{d}\ell(u) \cdot \mathrm{d}\ell(v)] \quad \text{for all } u, v \in T_P\mathcal{P}, \quad (10)$$

where $(\mathrm{d}\ell_x)(u)$ is the differential of $\ell_x \in C^\infty(\mathcal{P}, \mathbb{R})$ for all $P \in \mathcal{P}$. We will denote $\mathfrak{g}_P(u, v)$ by $\langle u, v\rangle$ whenever we think it is clear from the context that u, v are tangent vectors at $P \in \mathcal{P}$.

By means of Definition 1, $T_P\mathcal{P}$ defines an n-dimensional linear space over the reals equipped with an inner product given by the FIM. Since we want to be able to study covariance functions of possibly complex and/or vector-valued estimators we will use the language of tensor products which are defined, together with the so called complexification, in Appendix A.

From now on let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ denote the field of real or complex numbers. Using this notation we can treat the real and complex case in parallel. The FIM on $\mathbb{C} \otimes T_{P_\theta}\mathcal{P}$ is well defined by extending (10) to the complexification via (A.3).

Theorem 2 (Information Inequality) *Let $T : X \to \mathbb{K}$ be such that*

$$m : \Theta \to \mathbb{R}, \theta \mapsto E_{P_\theta}[T], \quad (11)$$

defines a smooth function $m \in C^\infty(\Theta, \mathbb{K})$. In particular T is an unbiased estimator of m. Then there holds, for all $\theta \in \Theta$,

$$\mathrm{var}_\theta(T) = E_{P_\theta}[|T - E_{P_\theta}T|^2] \geq \|\nabla m\|^2, \quad (12)$$

where the norm and the gradient ∇ correspond to the FIM on $\mathbb{K} \otimes T_{P_\theta}\mathcal{P}$.

Proof: We first check the claim for $\mathbb{K} = \mathbb{R}$. Let $\partial^i \in T_{P_\theta}\mathcal{P}$ denote the partial derivative w.r.t. θ_i acting on any smooth function of θ defined in a neighboorhood of $P_\theta \in \mathcal{P}$. Then $\partial^i\ell : x \mapsto \partial^i\ell_x$ is a statistic $X \to \mathbb{R}$ such that[4]

$$E_{P_\theta}[T \cdot \nabla\ell] = \nabla m,$$

and $E_{P_\theta}[\nabla\ell] = 0$ in particular. Now let $f : X \to \mathbb{K}$ be given by $f = T - m(\theta) - \langle\nabla m, \nabla\ell\rangle$ and observe that[5]

$$0 \leq E_{P_\theta}[f^2] = \mathrm{var}_\theta(T) - 2\langle\nabla m, \nabla m\rangle + \langle\nabla m, \nabla m\rangle.$$

[4]To see this note that

$$\begin{aligned} E_{P_\theta}[T \cdot \partial^i\ell] &= \int T(x)\frac{\partial}{\partial\theta_i}\log p_{(\cdot)}(x)\,\mathrm{d}P_\theta(x) \\ &= \int T(x)\left(p_\theta(x)^{-1}\frac{\partial}{\partial\theta_i}p_{(\cdot)}(x)\right)p_\theta(x)\mathrm{d}\nu(x) \\ &= \frac{\partial}{\partial\theta_i}\int T(x)p_\theta(x)\,\mathrm{d}\nu(x) = \frac{\partial}{\partial\theta_i}m. \end{aligned}$$

[5]This holds due to

$$\begin{aligned} E_{P_\theta}f^2 &= \mathrm{var}_\theta(T) - 2E_{P_\theta}(T - m(\theta))\langle\nabla m, \nabla\ell\rangle + E_{P_\theta}\langle\nabla m, \nabla\ell\rangle^2 \\ &= \mathrm{var}_\theta(T) - 2E_{P_\theta}[\langle\nabla m, T\nabla\ell\rangle] + E_{P_\theta}[(\mathrm{d}\ell(\nabla m) \cdot \mathrm{d}\ell(\nabla m)] \end{aligned}$$

and the fact that $E_{P_\theta}[\langle\nabla m, T\nabla\ell\rangle] = \langle\nabla m, E_\theta T\nabla\ell\rangle = \|\nabla m\|^2$.

For the complex valued case let $m = u + \jmath v$ with $u, v \in C^\infty(\Theta, \mathbb{R})$ and $T = R + \jmath S$ with $R, S : X \to \mathbb{R}$. Then

$$E_{P_\theta}[|T - m(\theta)|^2] = E_{P_\theta}[(R - u(\theta))^2] + E_{P_\theta}[(S - v(\theta))^2] \leq \|\nabla u\|^2 + \|\nabla v\|^2 = \|\nabla m\|^2,$$

which concludes the proof.[6] □

In applications one seldom tries to estimate scalar quantities. In Section III-C we treat the quite general case where one wants to estimate a function instead of a scalar. In the context of system identification this function corresponds to the transfer function of a LTI system.

C. The Covariance Function

Assume Θ is diffeomorphic to a sub-manifold $\mathcal{F}$ of a linear space $\mathcal{H}$ which consists of real- or complex-vector valued functions $F : \Omega \to \mathbb{K}^q$ defined on some set Ω. In other words the functions in $\mathcal{F}$ admit a parametrization defined on Θ. For example $\mathcal{H}$ could be $H^2(\Omega)$, the Hardy space of square integrable functions, analytic on the open unit disk $\Omega = \mathbb{D}$ or the open right half plane $\Omega = \{z \,|\, \mathrm{Re}(z) > 0\}$. After a reparametrization of $\mathcal{P}$ by $\{P_F\}_{F \in \mathcal{F}}$ we identify $T_{P_F}\mathcal{P}$ with $T_F\mathcal{F} \subseteq \mathcal{H}$ which is a space of functions. We are now ready to define the natural analog of a covariance matrix in the function space $\mathcal{F}$.

Definition 3 Given a statistic $T : X \to \mathcal{F}$ with $m(z) = E_{P_F}[T(z)]$ for some $m \in \mathcal{H}$ we define its *covariance at* $F \in \mathcal{F}$, denoted by $\mathrm{Cov}_F(T)(z, w)$, setting it equal to

$$E_{P_F}[(T(z) - m(z))(T(w) - m(w))^*], \tag{13}$$

for all $z, w \in \Omega$. Note that $\mathrm{Cov}_F(T)$ is defined on $\Omega \times \Omega$ and takes values in $\mathbb{K}^{q \times q}$. Also note that $\mathrm{Cov}_F(T) \geq 0$ in the sense of E.H. Moore, i.e., for all families $\{e_z\}_{z \in \Omega} \subseteq \mathbb{K}$ with finite support there holds[7]

$$\sum_{z,w \in \Omega} \bar{e}_z e_w \, v^* \mathrm{Cov}_F(T)(z, w) v \geq 0, \tag{14}$$

for all $v \in \mathbb{K}^q$.

To understand the interplay between covariance functions and the FIM we need a link between positive kernels, i.e., a function $K : \Omega \times \Omega \to \mathbb{K}^{q \times q}$ and inner product spaces $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ of $\mathbb{K}^q$-valued functions. This link is provided by the reproducing property of positive kernels – discussed in more detail in Appendix B. We choose to denote the estimator T in Theorem 4 by $\hat{F}$ to emphasize the fact that the Theorem assumes that T is an unbiased estimators of F.

[6]Note that we use $\nabla m = \nabla u - \jmath \nabla v$ which follows from $\mathrm{d}m = \mathrm{d}u + \jmath \mathrm{d}v$ and the definition of the gradient and the complexification of the inner product. More explicitely:

$$\begin{aligned} \mathrm{d}m(\Delta) &= \langle \Delta, \nabla u \rangle + \jmath \langle \Delta, \nabla v \rangle \\ &= \langle \Delta, \nabla u - j \nabla v \rangle = \langle \Delta, \nabla m \rangle , \end{aligned}$$

for all $\Delta \in T_{P_\theta}\mathcal{P}$.

[7]By finite support family we mean that $\{z \in \Omega \,|\, e_z \neq 0\}$ is a finite subset of Ω.

Theorem 4 (Abstract Cramér-Rao LB) *Let $\hat{F} : X \to \mathcal{F}$ denote an estimator with $E_{P_F}[\hat{F}(z)] = F(z)$ for all $z \in \Omega$ and all $F \in \mathcal{F}$. Then there holds in the sense of E.H. Moore*

$$\mathrm{Cov}_F(\hat{F}) \geq K_F \quad \textit{for all} \quad F \in \mathcal{F}, \tag{15}$$

where K_F is the reproducing kernel of $\mathbb{K}^{q \times q} \otimes T_{P_F}\mathcal{P}$ w.r.t. the Fisher Information metric.

Proof: Let $v \in \mathbb{K}^q$ and $C = \mathrm{Cov}_F(\hat{F})$. There holds that $\nabla \bar{e}_w v^* F(w) = e_w K_F(\cdot, w) v$ and thus by Theorem 2

$$\begin{aligned} \sum \bar{e}_z e_w v^* C(z, w) v &= E_{P_F}\left| \sum \bar{e}_w v^* (\hat{F}(w) - F(w)) \right|^2 \\ &\geq \left\| \sum e_w K_F(\cdot, w) v \right\|^2 \\ &= \sum \bar{e}_z e_w v^* K_F(z, w) v. \end{aligned}$$

holds for all $F \in \mathcal{F}$. □

D. Asymptotically Efficient Estimators

Up until now we just considered estimators where one sample $x \in X$ is observed. For time series analysis and system identification in particular it is of course of interest to study the properties of estimators $T = \{T_N \,|\, N \in \mathbb{N}\}$ which are given N samples in X, i.e., $T_N : X^N \to \mathcal{F}$, where on X^N we consider the product σ-algebra $\mathcal{X}^N$. Similarly $(X^\infty, \mathcal{X}^\infty)$ denotes the measurable sequence sample space. We assume $\mathcal{P}^\infty = \{P_F\}_{F \in \mathcal{F}}$ is an n-dimensional manifold parameterized by $\mathcal{F} \subseteq \mathcal{H}$. Moreover we assume there exists a number N_0 such that the restriction $P \mapsto P^N = P|_{\mathcal{X}^N}$ becomes injective on $\mathcal{P}^\infty$ for all $N \geq N_0$.

Definition 5 On $\mathcal{P}^\infty$ we define the *asymptotic (average) Fisher information metric* $\mathfrak{g}$ via

$$\mathfrak{g}_P(u, v) = \lim_{N \to \infty} \frac{\mathfrak{g}_{P^N}(u, v \,|N)}{N}, \tag{16}$$

for all $u, v \in T_P\mathcal{P}^\infty$ where $\mathfrak{g}(\cdot, \cdot \,|N)$ denotes the FIM on $\{P^N \,|\, P \in \mathcal{P}\}$. An estimator $\hat{F} = (\hat{F}_N)$ s.t. $\hat{F}_N : X^N \to \mathcal{F}$ is called *asymptotically efficient* if

$$N \cdot \mathrm{Cov}_F(\hat{F}_N) \to K_F \quad \text{as} \quad N \to \infty, \tag{17}$$

pointwise, where K_F is the reproducing kernel of $\mathbb{K}^{q \times q} \otimes T_{P_F}\mathcal{P}^\infty$ w.r.t. the asymptotic (average) FIM.

If $\hat{F}$ is an efficient estimator then $\mathrm{Cov}_F(\hat{F}_N) \approx N^{-1} K_F$ for a sufficiently large number N of samples. This is why the asymptotic FIM can be used for approximate variance and covariance quantification. It is well known that estimators which are asymptotically maximum likelihood (ML) are also asymptotically efficient [7].

IV. Consequences for System Identification

In Theorem 4 of Section III-C we derived the abstract CRLB for unbiased estimator in function spaces. In this Section we will interpret this result in the context of SYSID. Of special importance is the interpretation of the tangent space $T_F\mathcal{F}$ as a direct sum of subspaces of the Hardy space $H^2(\mathbb{T})$ discussed in Section IV-B. This interpretation allows us to compute the FIM and its reproducing kernel in practice

using real rational modules and Christoffel-Darboux type of identities.

A. Prediction Error Identification

Let $\mathcal{V} = \mathbb{R}^{\mathbb{N}}$, and let $O \subseteq H^2(\mathbb{T})$ denote the ring of stable proper rational functions [8]. We turn the sequence space $\mathcal{V}$ into an O-module by defining

$$O \times \mathcal{V}, (g, \xi) \mapsto L(g) * \xi \quad \text{with} \quad L(g) = (g_0, g_1, g_2, \cdots),$$

where $*$ denotes convolution and $g = \sum g_i z^{-i}$ is the expansion of g at infinity. Note that $z^{-1} \in O$ acts on $\mathcal{V}$ as a delay or right shift, i.e., $(z^{-1}\xi)(0) = 0$ and $(z^{-1}\xi)(t) = \xi(t-1)$ for all non-zero $t \in \mathbb{T}$. Hence the right shift is not surjective which is not a problem if one assumes zero initial conditions.

Let $\mathcal{V}_\sigma$ denote the product σ-algebra on $\mathcal{V}$ and $(\mathcal{V}, \mathcal{V}_\sigma, P)$ a probability space such that the projections $e_t : \xi \mapsto \xi_t$ are i.i.d. with zero mean and bounded variance $Ee_t^2 = \sigma^2$ for all $t \in \mathbb{T}$. We think of $e = (e_t)$ as a white noise sequence.

We seek to identify a system $F = (G, H) \in z^{-1}O \times O$ with H being a monic unit in O. Exciting it with by a known input sequence $u \in \mathcal{V}$ and observing its output y given by $y = Gu + He$.[8] The probability law of the random output sequence y depends on P, u and F. Since P is fixed and u is known it makes sense to denote this law by P_F; see Fig. 3.

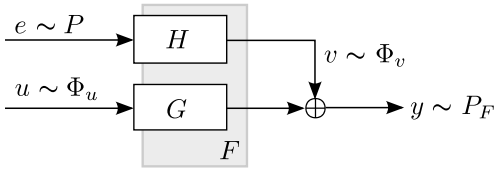

Fig. 3. Dependence of the law P_F of the output process y on $F = (G, H)$.

We assume the input $u \in \mathcal{V}$ is persistently existing such that $F \mapsto P_F$ is injective. Let $X^\infty = \mathcal{V} \times \mathcal{V}$ with product σ-algebra $\mathcal{X}^\infty$ such that every $x \in X^\infty$ represents a data sequence written as $x = (u_0, y_0, u_1, y_1, u_2, y_2, \dots)$. This data is assumed to be generated by a system $F \in \mathcal{F}$ where $\mathcal{F}$ is a finite dimensional manifold such that $(X^\infty, \mathcal{X}^\infty, \mathcal{P}^\infty)$ with $\mathcal{P}^\infty = \{P_F\}_{F \in \mathcal{F}}$ is a regular statistical model.

The Prediction Error framework is based on the asymptotically ML estimator sequence $\hat{F} = (\hat{F}_N)_{N=1,2,\dots}$ with $\hat{F}_N : X^N \to \mathcal{F}$ s.t. for all $x = (x_1, \dots, x_N)$ we have

$$\hat{F}_N(x) = \arg\min_{F \in \mathcal{F}} \frac{1}{N} \sum_{t=1}^{N} |[\hat{y}_F(x)]_t - y_t|^2, \qquad (18)$$

where N denotes the sample size and

$$\hat{y}_F(x) = H^{-1}Gu + (1 - H^{-1})y, \qquad (19)$$

the one step ahead predictor given $F = (G, H) \in \mathcal{F}$.

[8]Note that formally $y : \mathcal{V} \to \mathcal{V}$ is defined as the $\mathcal{F}$-measurable function which maps a potential noise realization $\xi \in \mathcal{V}$ to $Gu + H\xi \in \mathcal{V}$ which is the corresponding output sequence.

B. Computing Asymptotic FIM on the System Manifold

The n-dimensional system manifold $\mathcal{F}$ is a submanifold of the product manifold $\mathcal{G} \times \mathcal{H} \subseteq z^{-1}O \times O$ where $\mathcal{G}$ is the image of $\mathcal{F}$ under $(G, H) \mapsto G$ whereas $\mathcal{H}$ is the image under the corresponding complementary projection. If $\mathcal{F} = \mathcal{G} \times \mathcal{H}$ then G and H can be independently parameterized. In general for any $F = (G, H) \in \mathcal{F}$ we have the inclusion[9]

$$T_F\mathcal{F} \subseteq T_F(\mathcal{G} \times \mathcal{H}) \cong T_G\mathcal{G} \oplus T_H\mathcal{H}.$$

and thus every tangent vector of $\mathcal{F}$ at $F = (G, H)$ has two components one being tangent to $\mathcal{G}$ at G the other being tangent to $\mathcal{H}$ at H. The tangent vector at G is represented by a unique strictly proper rational function since $\mathcal{G} \subseteq z^{-1}O$. Similarly the tangent vector at H is represented by a unique proper rational function. In the following we interpret the abstract tangent vector $\partial_F \in T_F\mathcal{F}$ as element in $z^{-1}O \times O$ and denote this vector by $\partial_F = (\partial_G, \partial_H)$.

Since the PEM estimator $\hat{F}_N$ defined in (18) is asymptotically efficient for large N there holds

$$\mathrm{Cov}_F(\hat{F}_N) \approx N^{-1}K_F, \qquad (20)$$

where K_F is the reproducing kernel of $T_F\mathcal{F}$ w.r.t. the asymptotic FIM given by the integral expressions (22) in Theorem 6.

Theorem 6 *Let $F = (G, H) \in \mathcal{F}$ and $\mathfrak{g}$ denote the asymptotic FIM on $\mathcal{F} \cong \mathcal{P}^\infty$. Then there holds*

$$\mathfrak{g}_F(\partial_{F,i}, \partial_{F,j}) = \mathfrak{g}_F(\partial_{G,i}, \partial_{G,j}) + \mathfrak{g}_F(\partial_{H,i}, \partial_{H,j}). \qquad (21)$$

If we denote Φ_u and $\Phi_v = \sigma^2 HH^*$ the input and noise spectrum, respectively, then for $\partial_{F,i}, \partial_{F,j} \in T_F\mathcal{F}$ we have

$$\mathfrak{g}_F(\partial_{G,i}, \partial_{G,j}) = \int_{\mathbb{T}} (\partial_{G,i} \cdot \partial^*_{G,j})(\mathrm{e}^{\jmath\omega}) \, \frac{\Phi_u}{\Phi_v} \frac{\mathrm{d}\omega}{2\pi}, \qquad (22a)$$

$$\mathfrak{g}_F(\partial_{H,i}, \partial_{H,j}) = \int_{\mathbb{T}} (\partial_{H,i} \cdot \partial^*_{H,j})(\mathrm{e}^{\jmath\omega}) \, \frac{\sigma^2}{\Phi_v} \frac{\mathrm{d}\omega}{2\pi}, \qquad (22b)$$

where integrals are taken over the unit circle $\mathbb{T}$ w.r.t. the standard Lebesgue measure. Note that Φ_v varies with F on $\mathcal{F}$ since it depends on H.

Proof: Is an immediate consequence of Lemma A.4 and Theorem A.5. □

Remark 7 The pair (Φ_u, Φ_v) determines $\mathfrak{g}_F$ but not the other way around. To see this, let X^{q_G} denote the smallest real rational module containing $T_G\mathcal{G}$ (see Appendix D for the Definition of X^q). Then the moments $\mu_k \in \mathbb{R}$ with

$$\mu_k = \frac{1}{2\pi} \int_{\mathbb{T}} z^k \, \frac{\Phi_u}{\Phi_v} \, \frac{\mathrm{d}\omega}{|q_G|^2} \quad k = 0, \dots, \deg(q_G) - 1, \qquad (23)$$

determine $\mathfrak{g}_{F|T_G\mathcal{G}}$ since $\frac{z^k}{q_G}$ form a basis for X^{q_G}; see Fig. 4.

[9]If $T_F\mathcal{F}$ equals the sum $T_G\mathcal{G} \oplus T_H\mathcal{H}$ this means that the transfer functions G and H can be independently parametrized. For example this is the case when $\mathcal{F}$ is given by a BJ model structure. In general we have no equality; pick, e.g., $\mathcal{F}$ from the ARX model class.

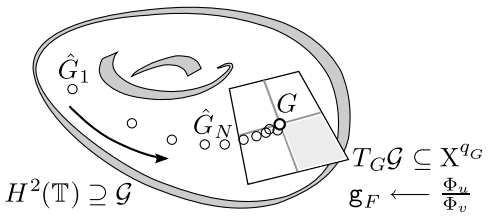

Fig. 4. The Fisher Information metric $\mathfrak{g}_F$ on the tangent space of $\mathcal{G}$ at G can be expressed in terms of a $\frac{\Phi_u}{\Phi_v}$ weighted $H^2(\mathbb{T})$ metric on the real rational module X^{q_G}.

Similarly $\mathfrak{g}_{F|T_H\mathcal{H}}$ is determined by

$$\mu_k = \frac{1}{2\pi}\int_{\mathbb{T}} z^k \frac{\sigma^2}{\Phi_v}\frac{\mathrm{d}\omega}{|q_H|^2} \quad k = 0, \ldots, \deg(q_H) - 1, \quad (24)$$

where X^{q_H} is the real rational module hull of $T_H\mathcal{H}$.

C. Computing Asymptotic Covariance from FIM

In general the computation of the reproducing kernel K_F of $T_F\mathcal{F}$ w.r.t. the asymptotic average FIM amounts to finding an orthonormal basis (ONB), i.e., inverting a Gram-matrix of a general basis of $T_F\mathcal{F}$ using the Aitken-Collar-Berg Lemma A.3. However simplifications can occur:

1) If $T_F\mathcal{F} = T_G\mathcal{G} \oplus T_H\mathcal{H}$ then
$$K_F = \begin{bmatrix} K_{F|\mathcal{G}} & \\ & K_{F|\mathcal{H}} \end{bmatrix}. \quad (25)$$
2) For $\frac{\Phi_u}{\Phi_v} = |\rho|^2$ and $\rho T_G\mathcal{G} = \mathrm{X}^q$ for some polynomial $q \in \mathbb{R}[z]$ then $K_{F|\mathcal{G}} = K$ with K given by (A.12), see e.g., [3][(27,49,50)] where X^q is X_n and $\rho = S_\zeta/H$.
3) If $T_G\mathcal{G} = \mathrm{X}^q$ then $K_{F|\mathcal{G}}$ can be computed from K defined in (A.12) via
$$K_{F|\mathcal{G}}(z,w) = \int K(\mathrm{e}^{\jmath\omega}, w)\cdot K^*(\mathrm{e}^{\jmath\omega}, z)\ \tfrac{\Phi_v}{\Phi_u}\tfrac{\mathrm{d}\omega}{2\pi}, \quad (26)$$
where the integral is weighted with the inverse signal-to-noise (SNR) ratio.

One should note that if $T_F\mathcal{F}$ is a proper subspace of $T_G\mathcal{G} \oplus T_H\mathcal{H}$, which is the case for model classes where G and H cannot be independently parametrized, then $K_F\mathcal{F}$ is no longer diagonal. There holds

$$K_{F|\mathcal{F}} = \mathfrak{g}_F(\Pi K_{F|\mathcal{G}\times\mathcal{H},w}, K_{F|\mathcal{G}\times\mathcal{H},z}), \quad (27)$$

with Π denoting the orthogonal projection onto $T_F\mathcal{F}$ w.r.t. the FIM metric $\mathfrak{g}_F$.

V. Conclusions

In this paper we have established that the auto-covariance of an unbiased function estimator is a positive kernel which can be bounded from below by the reproducing kernel of the tangent space of the function manifold in a general statistical space framework. This bound becomes asymptotically tight if the function estimator is efficient. The consequence these results have for system identification is that the problem of quantifying the auto-covariance of a transfer function estimator splits into two subproblems: determining the tangent space of the model manifold at the system which generated the data, and computing its reproducing kernel with respect to the SNR-weighted inner product. Often the tangent space forms a real rational module; in any case there exists a smallest real rational module containing it; this fact gives rise to closed form expressions for the reproducing kernel which can be used in optimal-input design. For future work we want to relax the system in the model manifold assumption and allow the data generating system to be in the topological closure of the model manifold but not in the manifold itself also referred to as overmodelling. In such situations quantifying the covariance matrix of a paramater vector is an ill-defined problem. We believe that by means of the coordinate free methods developed in this paper autocovariance quantification with overmodelling will be possible even if the topological closure is not a manifold.

VI. Acknowledgements

The authors would like to thank P.-A. Absil and B.D.O. Anderson for helpful discussions.

Appendix

A. Tensor Products and Complexification

Tensor products form the abstraction of Kronecker products for matrices. Recall that if $A \in \mathbb{R}^{a\times c}$ and $B \in \mathbb{R}^{b\times d}$ the Kronecker product $A \otimes B \in \mathbb{R}^{ab\times cd}$ is a block matrix with block size $b \times d$ whose i,j-th block given by $a_{ij}B$.[10] The main disadvantage of Kronecker products is that they are not applicable to general linear spaces before choosing a basis. For instance choosing a basis for the tangent space in Theorem 2 amounts to the choice of a specific parametrization. If the tangent space is a space of functions where inner products can be computed in a natural way, e.g., say on the unit circle, it is of interest to preserve the structure instead of choosing an arbitrary basis treating this space as if it were a general inner product space. This motivates the following abstract Definition of Tensor products.

Definition A.1 Let U, V denote two linear spaces over the real numbers of dimension n and m respectively. A linear space $(W, \otimes)$ of dimension nm equipped with a bilinear map

$$\otimes : U \times V \to W, (u,v) \mapsto u \otimes v,$$

is called *tensor product* of U and Y if $\{u_i \otimes v_j\}$ is a basis of W whenever $\{u_i\}$ is a basis of U and $\{v_j\}$ is a basis of V. If this is the case W is denoted by $W = U \otimes V$.

Moreover if $\langle\cdot,\cdot\rangle_U$ and $\langle\cdot,\cdot\rangle_V$ denote inner products on U and V respectively there exists a unique inner product $\langle\cdot,\cdot\rangle$ on $U \otimes V$ such that for all $u_i \in U$, $v_i \in V$

$$\langle u_1 \otimes v_1, u_2 \otimes v_2\rangle := \langle u_1, u_2\rangle_U \cdot \langle v_1, v_2\rangle_V . \quad (A.1)$$

If $n = 2q$ and we have a $\mathbb{C}$-linear space structure on U then $W = U \otimes V$ becomes a $\mathbb{C}$-linear space as well by defining the $\mathbb{R}$-bilinear map $\odot : \mathbb{C} \times W \to W$ via

$$z(u \otimes v) := (zu) \otimes v \quad \text{for all } z \in \mathbb{C}, \quad (A.2)$$

[10] This example is treated by Definition A.1 below if one sets $U = \mathbb{R}^{a\times c}$ and $V = \mathbb{R}^{b\times d}$. One can gain insight by looking at the special case: $c = 1$ and $b = 1$. Note that in this $A \otimes B$ are rank 1 matrices which span $\mathbb{R}^{a\times b}$.

and all $u \in U$, $v \in V$. Note that $U \otimes V$ has $\mathbb{C}$-dimension $q \cdot m$. Finally if U has a complex inner product $\langle\!\langle \cdot, \cdot \rangle\!\rangle_U$ then

$$\langle\!\langle u_1 \otimes v_1, u_2 \otimes v_2 \rangle\!\rangle := \langle\!\langle u_1, u_2 \rangle\!\rangle_U \langle v_1, v_2 \rangle_V , \qquad \text{(A.3)}$$

can be uniquely extended to a complex inner product on $U \otimes V$.

Definition A.2 A special case of this construction is $U = \mathbb{C}$ with inner product given by $\langle\!\langle u_1, u_2 \rangle\!\rangle_U = u_2^* u_1$. We call $\mathbb{C} \otimes V$ the *complexification* of V. The complexification $\mathbb{C} \otimes V$ is a linear space over $\mathbb{C}$ whose $\mathbb{C}$-dimension equals the $\mathbb{R}$-dimension of V. Instead of writing $z \otimes v$ one writes $\alpha v + \jmath \beta v$ with $\alpha, \beta \in \mathbb{R}$ such that $z = \alpha + \jmath\beta$. With this notation $V, \jmath V \subseteq \mathbb{C} \otimes V$ are $\mathbb{R}$-subspaces such that

$$\mathbb{C} \otimes V = V \oplus \jmath V, \qquad \text{(A.4)}$$

where the sum is orthogonal w.r.t. to the real inner product $\langle \cdot, \cdot \rangle$ defined in (A.1) with $\langle 1, \jmath \rangle_U = 0$ and $1, \jmath \in U$ having unit norm. For $V = \mathbb{R}^n$ we have $\mathbb{C}^n = \mathbb{C} \otimes V$. Also note that trivially $V = \mathbb{R} \otimes V$.

B. Positive Kernels Reproducing Spaces

For any space $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ of $\mathbb{K}^q$-valued functions on Ω the associated kernel $K_{\mathcal{F}}$ is defined for all $\alpha \in \mathbb{K}^q$, $z \in \Omega$ by $K_{\mathcal{F}}(z, w)\alpha = K_{w,\alpha}(z)$ where $K_{w,\alpha} \in \mathcal{F}$ is the unique Riesz-representantive of the evaluation $\mathcal{F} \to \mathbb{K}, f \mapsto \alpha^* f(w) = \langle f, K_{w,\alpha} \rangle$ which is assumed to be continuous [2]. If $\mathcal{F}$ is finitely generated, then

$$K_{\mathcal{F}}(z, w) = \sum_{i=1}^{n} b_i(z) b_i^*(w), \qquad \text{(A.5)}$$

where $\{b_i\}_{i=1}^n$ can be any ONB of $\mathcal{F}$. Conversely any positive kernel K generates a space $\mathcal{F}_K$ which is the completion of the linear space generated by $\{K_{w,\alpha}\}_{w \in \Omega, \alpha \in \mathbb{K}^q}$ w.r.t. the inner product defined by

$$\langle K_{w,\alpha}, K_{z,\beta} \rangle = \beta^* K(z, w) \alpha, \qquad \text{(A.6)}$$

for all $z, w \in \Omega$ and $\alpha, \beta \in \mathbb{K}^q$. An easy to check yet fundamental property is that $K \mapsto \mathcal{F}_K$ and $\mathcal{F} \mapsto K_{\mathcal{F}}$ are inverse to each other.

Lemma A.3 (ABC) *Let $g = [g_1, \ldots, g_n]$ denote a basis for the $\mathbb{K}$-inner product space $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ consisting of $\mathbb{K}$-valued functions. Moreover let $G_{ij} = \langle g_j, g_i \rangle$ denote the Gramian matrix. Then $g(z) G^{-1} g^*(w)$ is the reproducing kernel of $\mathcal{F}$.*

C. Asymptotic Efficiency of PEM

The proof of Lemma A.4 follows from direct computation and is therefore omitted.

Lemma A.4 *For fixed $x \in X^\infty$ the one step-ahead predictor defined in (19) can be viewed as a function*

$$\hat{y}(x) : \mathcal{F} \to \mathbb{R}^{\mathbb{N}}, F \mapsto \hat{y}_F(x), \qquad \text{(A.7)}$$

which has smooth components $[\hat{y}(x)]_t : \mathcal{F} \to \mathbb{R}$. In particular, given any $\partial_F \in T_F\mathcal{F}$, the sequence $(\partial_F [\hat{y}(x)]_t)_{t \geq 1}$, denoted by $\partial_F \hat{y}(x)$, is well defined and can be expressed in terms of $x = (u, y)$ and $F = (G, H)$ via

$$\partial_F \hat{y}(x) = H^{-1}(\partial_G u + \sigma \partial_H e), \qquad \text{(A.8)}$$

where e is such that $y = Gu + He$ holds.

The following Theorem A.5 was proven in Caines and Ljung [9] and is of fundamental importance. It says that the PEM is asymptotically efficient and it gives a way to compute the asymptotic (average) FIM.

Theorem A.5 *Define $\hat{g}_F(\partial_{F,i}, \partial_{F,j} | N) : X^\infty \to \mathbb{R}$ via*

$$x \mapsto (N\sigma)^{-2} \cdot \textstyle\sum_{t=1}^{N} [\partial_{F,i} \hat{y}(x)]_t \cdot [\partial_{F,j} \hat{y}(x)]_t, \qquad \text{(A.9)}$$

for all $\partial_{F,i}, \partial_{F,j} \in T_F\mathcal{F}$. Then as $N \to \infty$ we have

$$\hat{g}_F(\partial_{F,i}, \partial_{F,j} | N) \to g_F(\partial_{F,i}, \partial_{F,j}) \quad (P_F\text{-a.s.}), \qquad \text{(A.10)}$$

where g denotes the asymptotic FIM on $\mathcal{F} \cong \mathcal{P}^\infty$.

D. Real Rational Modules as Coinvariant Spaces

A very simple and yet fundamental class of subspaces in $H^2(\mathbb{T})$ given by real rational modules also called coinvariant subspaces [10].

Definition A.6 Let $q \in \mathbb{R}[z]$ denote a stable polynomial. The associated *real rational module* is denoted by X^q with

$$\mathrm{X}^q = \left\{ \frac{p}{q} \in \mathbb{R}(z),\, p \in \mathbb{R}(z), \deg(p) < \deg(q) \right\}. \qquad \text{(A.11)}$$

The reproducing kernel of $\mathbb{C} \otimes \mathrm{X}^q \subseteq H^2(\mathbb{T})$ is given by the Christoffel-Darboux formula

$$K(z, w) = \frac{1}{q(z) q^\star(w)} \cdot \frac{q^\star(z) q(w) - q(z) q^\star(w)}{1 - z\bar{w}}, \qquad \text{(A.12)}$$

with $q^\star(z) = z^n q(z^{-1})$ and $n = \deg(q)$.

REFERENCES

[1] J. M. Oller and J. M. Corcuera, "Intrinsic analysis of statistical estimation," *The Annals of Statistics*, vol. 23, no. 5, pp. 1562–1581, October 1995.

[2] C. Carmeli, E. D. Vito, and A. Toigo, "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem," *Analysis and Applications*, 2006.

[3] B. Ninness and H. Hjalmarsson, "Variance error quantifications that are exact for finite-model order," in *Transactions on Automatic Control*, vol. 49, no. 8, August 2004, pp. 1275–1290.

[4] T. Ivanov, P.-A. Absil, B. Anderson, and M. Gevers, "Application of real rational modules in system identification," in *47th IEEE Conference on Decision and Control*, 2008.

[5] U. Helmke and P. A. Fuhrmann, "Tangent spaces of rational matrix functions," *Linear Algebra and its Applications*, vol. 271, pp. 1–40, 1998.

[6] D. Alpay, L. Baratchart, and A. Gombani, *The Differential Structure Of Matrix-valued Rational Inner Functions*. Birkhlauser Verlag,, 1994, vol. Nonselfadjoint Operators and Related Topics.

[7] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry (Translations of Mathematical Monographs)*. American Mathematical Society, April 2001.

[8] M. Vidyasagar, *Control System Synthesis A Factorization Approach*. MIT Press Cambridge, 1985.

[9] P. Caines and L. Ljung, "Asymptotic Normality of Prediction Error Estimators for Approximate System Models," in *Stochastics*, vol. 3, January 1979, pp. 29–46.

[10] P. A. Fuhrmann, *A Polynomial Approach to Linear Algebra*. Springer-Verlag New York, Inc., 1996.