The Information Inequality on Function Spaces given a Singular Information Matrix†

Tzvetan Ivanov*, P.-A. Absil* and Michel Gevers*

***Abstract*— In this work we extend the scope of the classical Cramér-Rao lower bound, or information inequality, from Euclidean to function spaces. In other words we derive a tight lower bound on the autocovariance function of a function estimator. We do this in the context of system identification. Two key elements of system identification are experiment design and model selection. The novel information inequality on function spaces is important for model selection because it allows the user to compare estimators using different model structures. We provide a consistent treatment of the case where the Fisher information matrix is singular. This makes it possible to take into account that in optimal experiment design one tries to mask those parts of the system non-identifiable, which are irrelevant for the application.**

Keywords: System Identification, Autocovariance, Fisher Information Geometry, Reproducing Kernel, Duality

I. Introduction

Parameter estimation techniques based on information inequalities like the Cramér-Rao lower bound have found a huge amount of applications in statistics, system identification and machine learning [11]. These bounds can be used to quantify the covariance matrix of asymptotically efficient parameter estimators as the sample size tends to infinity [12]. Various optimal experiment design techniques have been proposed to reduce parametric uncertainty. Traditionally optimality criteria such as A-,D- and E-optimality serve that purpose by maximizing the trace, determinant or minimal eigenvalue of the information matrix respectively. However, in system identification the primary goal is not to estimate a parameter but to estimate a system. That this is indeed an issue becomes clear if one keeps in mind that there exist a variety of natural parametrizations for a given model structure. For instance the model structure consisting of transfer function P with

$$P(z) = \frac{\theta_1 z^{-1} + \cdots + \theta_n z^{-n}}{1 + \theta_{n+1} z^{-1} + \cdots + \theta_{2n} z^{-n}} = \sum\nolimits_{i=1}^{\infty} g_i z^{-i},$$

can be parameterized using $\theta = (\theta_1, \ldots, \theta_{2n})$ or the Markov parameters $g = (g_1, \ldots, g_{2n})$.

†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, P.-A. Absil 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,absil,gevers}

If, for example, we maximize the trace of the information matrix in the g-coordinate system and in the θ-coordinate system we perform

$$\min \mathbf{E}\|\hat{g} - g\|^2 \quad \text{and} \quad \min \mathbf{E}\|\hat{\theta} - \theta\|^2,$$

respectively, where $\hat{\cdot}$ denotes the estimator. For some application mean-square error (MSE) of $\hat{g}$ might be important, for another the MSE of $\hat{\theta}$, and for yet another the MSE of a completely different parameter vector might be of interest [2], [13]. In applications, such as robust control, it is more natural to write a specification in terms of the function estimator. In order to remain in the computationally feasible realm of convex optimization such performance specifications have to be linked to a 2-norm which in a statistical framework is the variability metric. The autocovariance function $\Phi : \mathbb{T} \times \mathbb{T} \to \mathbb{C}$ given by

$$\Phi(z, w) := \mathbf{E}[\Delta(z)\Delta(w)^*] \quad \text{with} \quad \Delta = P - \hat{P},$$

and $\mathbb{T} \subseteq \mathbb{C}$ denoting the unit circle provides this link. In other words it is possible to derive sufficient conditions for natural performance specifications based on the function mismatch Δ by using the autocovariance function.

In this article will derive a novel information inequality which generalizes the classical Cramér-Rao lower-bound, of the form $C \geq J^{-1}$, in such a way that it is more suitable for function estimation. In the new inequality the covariance matrix C is replaced by the autocovariance function Φ, and the inverse information matrix J^{-1} by the reproducing kernel of the information metric. We will derive all our results without making the assumption that J is non-singular.

The article is structured as follows: In Section II we use a simple example to illustrate that a biased estimator, for which the true system is not even an element of the model structure, can outperform an unbiased estimator. In Section III we use duality theory to describe the inverse relationship between inner-products and reproducing kernels. In this context the Fisher information metric refers to the inner-product whose Gramian matrix is the classical Fisher information matrix. In Section IV we introduce the concept of compression for inner products. This concept is used in Section V to derive the information inequality for function spaces given a possibly singular Fisher information matrix. In Section VI we conclude.

Notation: $\mathbb{K}^*$ denotes the field of real or complex numbers. If $\mathbb{K} = \mathbb{C}$, and $k \in \mathbb{K}$, then k^* denotes the complex conjugate. If $\mathbb{K} = \mathbb{R}$, then $k^* = k$.

II. Motivational Example

Let $\hat{P}_N$ denote the standard LS-estimator of the transfer function P given input-output measurements $(x_1, \dots, x_N)$ with $x_t = (u_t, y_t)$ in the following statistical model structure

$$y = Pu + e \quad \text{with} \quad P(z) = \sum_{i=1}^{30} \theta_i z^{-i}, \tag{1}$$

where e denotes Gaussian white noise with variance σ_e^2.

We shall compare $\hat{P}_N$ with the reduced order estimator $\hat{R}_N$, defined by $\hat{R}_N = \rho(\hat{P}_N)$, where ρ is a model reduction map. As a concrete example consider the reduced order model structure given by

$$\mathcal{R} = \left\{ \frac{r_1}{r_2 + z} \,\middle|\, r \in \mathbb{R}^2 \right\}. \tag{2}$$

Since $R \in \mathcal{R}$ is determined by its value at the interpolation point $z_0 := \mathrm{e}^{\jmath\pi/2} = \jmath$, we may, instead of r, use

$$\tilde{r} = (\tilde{r}_1, \tilde{r}_2) = (\operatorname{Re} R(z_0), \operatorname{Im} R(z_0)), \tag{3}$$

as a coordinate vector.[1] The model reduction map

$$\rho : \mathcal{P} \to \mathcal{R}, \quad P \mapsto R \quad \text{with} \quad R(z_0) = P(z_0), \tag{4}$$

is thus well defined and smooth. For the low order estimator $\hat{R}_N$ we shall use a discrete input spectrum with input power σ_u^2 distributed evenly across $-\pi/2$ and $\pi/2$. For the high order estimator $\hat{P}_N$ we use the same input power but with a constant input spectrum, which corresponds to white inputs.

The simulation results, see Fig. 1, indicate that the reduced order estimator can outperform the full order estimator for high frequencies, despite the inherent bias $\rho(P) - P$ of the reduced order estimator. This is due to the difference of the corresponding variance functions, see Fig. 2.

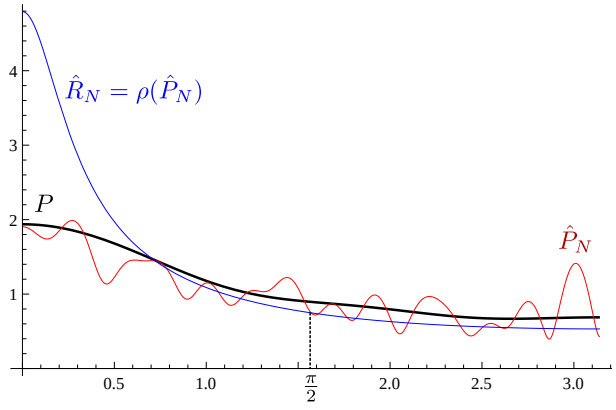

Fig. 1. Absolute value curves of the true system $P(\mathrm{e}^{\jmath\omega})$, a realization of the full order estimator $\hat{P}_N(\mathrm{e}^{\jmath\omega})$, and the corresponding reduced order estimator $\hat{R}_N = \rho(\hat{P}_N)$, over the frequency band $[0, \pi]$. The reduced order estimator outperforms the full order estimator on the high frequency band $[\pi/2, \pi]$. For low frequencies the full order estimator performs better. The simulation was performed with $N = 100$ and $\sigma_e^2/(N\sigma_u^2) = 0.05$

[1] A simple calculation reveals that the correspondence between $\tilde{r}$ and r is given by $r = \begin{bmatrix} -1 & \tilde{r}_1 \\ 0 & \tilde{r}_2 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \tilde{r}_2 \\ -\tilde{r}_1 \end{bmatrix}$.

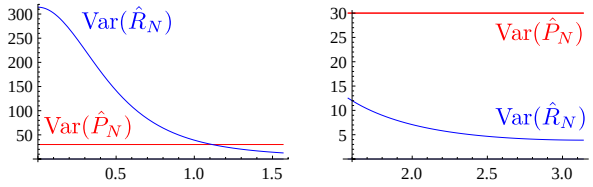

Fig. 2. Variance functions of the full order estimator $\hat{P}_N$ and the reduced order estimator $\hat{R}_N$ normalized by $N \cdot \sigma_u^2/\sigma_e^2$. For high frequencies the reduced order estimator outperforms the full order estimator. For low frequencies the order is reversed.

For optimal model selection it is therefore of interest to quantify the variance function of $\hat{R}_N$ as well as the variance of arbitrary functionals of $\hat{R}_N$. The preceeding example shows that the autocovariance function of $\hat{R}_N$ is finite despite the fact that the system was not in the model set, i.e., $P \notin \mathcal{R}$, and despite the fact that the the input, with a discrete power spectrum supported on $\{-\pi/2, \pi/2\}$, results in a singular Fisher information matrix on $\mathcal{P}$. The tools presented in the next sections will allow us to treat this problem.

III. Preliminaries on Inner-Products

The Fisher-information matrix J is, technically speaking, the Gramian matrix of an inner-product. In the classical CRLB one computes J^{-1}. Naturally one would expect that in a coordinate free framework this must correspond to the inversion of an inner-product. However, given an inner-product $\langle \cdot, \cdot \rangle : U \times U \to \mathbb{K}$ it is not really clear what it should mean to invert it. In the next section we shall see that we can view an inner-product as a map from the space U into its conjugate-dual space U^*. From this perspective, inversion is conceptually straightforward and yields and inner-product on U^*. We also show that the computation of reproducing kernels can be seen as a special case of this procedure.

A. Setup and Notation

To simplify notation introduce the conventions:

1) U is a *linear space* $:\Leftrightarrow$ finite dimensional and closed under linear combinations with scalars in $\mathbb{K}$.
2) T is a *linear map* $:\Leftrightarrow$ it is additive and commutes with multiplication by scalars in $\mathbb{K}$.
3) $\langle \cdot, \cdot \rangle$ is an *inner-product* on $U :\Leftrightarrow \mathbb{K}$-valued, positive-definite and satisfies $\langle ku, v \rangle = k \langle u, v \rangle = \langle u, k^* v \rangle$.

B. Inner-products and their associated operators

By property 3) above, i.e., by conjugate symmetry, fixing the first component of the inner-product yields a conjugate-linear functional on U. This makes Definition 1 natural.

Definition 1 The space of all conjugate-linear functionals on U is denoted by U^*. The linear map

$$G : U \to U^* \quad \text{with} \quad (Gu)(v) = \langle u, v \rangle \quad \forall u, v \in U, \tag{5}$$

is called the *associated operator* of the inner-product.

Every inner-product $\langle \cdot, \cdot \rangle$ corresponds to a unique operator $G : U \to U^*$. The converse is true if and only G is positive-definite in the sense of Definition 2.

Definition 2 The *adjoint* of linear map $G : U \to U^*$ is again a map from U to U^* and defined by $(\tilde{G}u)(v) := (Gv)(u)^*$ for all $u, v \in U$. An operator with $G = \tilde{G}$ is called *self-adjoint*. If in addition $(Gu)(u) \geq 0$ holds for all $u \in U$ one calls G *positive-semidefinite* which is denoted by $G \geq 0$. A non-singular $G \geq 0$ is denoted by $G > 0$ and called *positive-definite*. Similarly one writes $G \geq H$ or $H \leq G$ for two self-adjoint operators $G, H : U \to U^*$ satisfying $G - H \geq 0$.

These definitions ensure, in a coordinate-free manner, a bijective correspondence, which relates inner-products to the cone of positive-definite operators in the space of self-adjoint operators. We therefore shall, for the rest of this paper, treat the term inner-product as a synonym for the associated positive definite operator.

C. Duality and Reproducing Kernels

In Theorem 4 we shall see that an inner-product on a linear space induces an inner-product on its dual. In Theorem 10 we shall see that if the original space is a function space the induced inner-product on the dual admits an interpretation as a reproducing kernel. Before we can state these theorems we need to define the notion of a dual map.

Definition 3 The *dual map* of a map $T : U \to V$ is

$$T^* : V^* \to U^* \quad \text{with} \quad (T^*\ell)(u) = \ell(Tu) \tag{6}$$

for all $\ell \in V^*$ and $u \in U$.

Theorem 4 *Let U denote a linear space with an inner-product $G : U \to U^*$. For $G^{-*} := (G^{-1})^*$ we obtain*

$$(G^{-*}\ell)(\ell) = \sup\{|\ell(u)|^2 \,|\, (Gu)(u) = 1\}, \tag{7}$$

for all $\ell \in U^*$. In particular $G^{-*} : U^* \to U^{**}$ corresponds to an inner-product on U^*.

Proof: Since G is non-singular for every $\ell \in U^*$ there exists a unique $v \in U$ such that $Gv = \ell$. We can thus express the left-hand side of (7) by

$$(G^{-*}\ell)(\ell) = \ell(G^{-1}\ell) = (Gv)(G^{-1}Gv) = (Gv)(v).$$

Since G corresponds to an inner-product we can apply the Cauchy-Bunyakovsky-Schwarz inequality (CBS) to check that

$$|\ell(u)|^2 = |(Gv)(u)|^2 \leq (Gv)(v), \tag{8}$$

holds for all $u \in U$ with $(Gu)(u) = 1$.

Moreover the CBS (8) becomes an equality for $u = v$. This proves that the right hand side of (7) equals the left hand side. It remains to check that G^{-*} is positive-definite. To check that $\tilde{G}^{-*} = G^{-*}$, (i.e., G^{-*} is self-adjoint) let $\ell, \eta \in U^*$ with $\ell = Gu$ and $\eta = Gv$ and calculate

$$(G^{-*}\eta)(\ell) = (Gv)(u) = (Gu)(v)^* = (G^{-*}\ell)(\eta)^*,$$

where we used $G = \tilde{G}$. Positive-definiteness follows from (7) and the fact that G is non-singular. This proves that G^{-*} corresponds to an inner-product on U^*. □

Remark 5 In the proof of Theorem 4 we showed that

$$(G^{-*}\ell)(\ell) = (Gv)(v),$$

if $v \in V$ is such that $\ell = Gv$. This fact is useful because it relates properties of G^{-*} to properties of G.

Definition 6 For $G : U \to U^*$ self-adjoint we define the *induced quadratic forms* via

$$G_Q(u) := (Gu)(u) \quad \text{and} \quad G_Q^{-*}(\ell) := (G^{-*}\ell)(\ell), \tag{9}$$

for all $u \in U$ and $\ell \in U^*$ respectively.

Remark 7 It is an important fact that any self-adjoint operator is uniquely determined by the quadratic form which it induces. For inner-product spaces this result is referred to as polarization identity [4].

Definition 8 Let U denote a linear space. Given a function $\eta : U \to \mathbb{K}$ we define $\bar{\eta}$ via $\bar{\eta}(u) = \eta(u)^*$. We call

$$\mathrm{ev}|_\Omega := \{\mathrm{ev}_z : U \to \mathbb{K} \,|\, z \in \Omega\}, \tag{10}$$

an *evaluation structure* on U if $\overline{\mathrm{ev}}_z \in U^*$ for all $z \in \Omega$.

One says that the pair $(U, \mathrm{ev}|_\Omega)$ forms a *linear function space* if, given that $\mathrm{ev}_z(u)$ vanishes for all points z in Ω, one can conclude that $u = 0$.

Definition 9 Let $(U, \mathrm{ev}|_\Omega)$ denote a linear function space and G denote an inner-product on U. The function:

$$K : \Omega \times \Omega \to \mathbb{K} \quad \text{with} \quad K(z, w) = (G^{-*}\overline{\mathrm{ev}}_w)(\overline{\mathrm{ev}}_z), \tag{11}$$

is called *the reproducing kernel of $(U, \mathrm{ev}|_\Omega)$ w.r.t. G. This agrees with the classical definitions in analysis [14].*

In linear function spaces it is common to identify vectors $u \in U$ with functions $\Omega \to \mathbb{K}, z \mapsto \mathrm{ev}_z(u)$ if the evaluation structure is clear. In particular one writes $u(z)$ instead of $\mathrm{ev}_z(u)$. We follow this tradition in Theorem 10.

Theorem 10 *Let $(U, \mathrm{ev}|_\Omega)$ denote a linear function space and G denote an inner-product on U. For all $w \in \Omega$ we define $K_w = \tilde{G}^{-1}\overline{\mathrm{ev}}_w$. Then:*

1) *$K_w \in U$ with $K_w(z) = K(z, w)$ for all $z, w \in \Omega$,*
2) *$(Gu)(K_w) = u(w)$ for all $w \in \Omega$ and $u \in U$,*

where K denotes the reproducing kernel $(U, \mathrm{ev}|_\Omega)$ w.r.t. G.

Proof: For any two points $z, w \in \Omega$ we have

$$\begin{aligned} K(z, w) &= (G^{-*}\overline{\mathrm{ev}}_w)(\overline{\mathrm{ev}}_z) \\ &= \overline{\mathrm{ev}}_w(G^{-1}\overline{\mathrm{ev}}_z) \\ &= \mathrm{ev}_z(\tilde{G}^{-1}\overline{\mathrm{ev}}_w)^* = K_w(z), \end{aligned}$$

which proves property 1). For all $w \in \Omega$ and $u \in U$ we have

$$\begin{aligned} (Gu)(K_w) &= (Gu)(\tilde{G}^{-1}\overline{\mathrm{ev}}_w) \\ &= (GG^{-1}\overline{\mathrm{ev}}_w)(u)^* \\ &= \overline{\mathrm{ev}}_w(u)^* = u(w), \end{aligned}$$

and hence property 2) is verified. □

Finally we note that, just like systems can be described by different parameter vectors, systems can also be described using different functions. In order to allow this type of flexibility we derived our results in an abstract framework where the function space structure enters via evaluations.

Example 1 Let $f \in \mathbb{R}(z)$ denote a Schur-stable strictly proper rational function. We can evaluate f in the frequency domain, by defining $\mathrm{ev}_{|(-\pi,\pi]}$ via

$$\mathrm{ev}_\omega(f) := f(\mathrm{e}^{\jmath\omega}) \quad \text{for all} \quad \omega \in (-\pi, \pi], \tag{12}$$

or in the time domain $\mathbb{N} = \{1, 2, \dots\}$, by defining $\mathrm{ev}_{|\mathbb{N}}$ via

$$\mathrm{ev}_t(f) := \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\mathrm{e}^{\jmath\omega})\mathrm{e}^{-\jmath t\omega}\,\mathrm{d}\omega \quad \text{for all} \quad t \in \mathbb{N}. \tag{13}$$

IV. Compression of Inner-Products

In order to state the information inequality given a possibly singular information metric, we put forward the concept of compressing an inner-product by a linear map. Those interested only in the application of the results to information geometry can skip to Section V after reading Theorem 11 and Definition 12.

Theorem 11 *Let U be a linear space and $G : U \to U^*$ positive-semidefinite. For any subjective linear map $T : U \to V$ there exists a unique $F_{\max} \geq 0$ corresponding to an inner-product on V such that for all $F : V \to V^*$ the following equivalence holds*

$$T^*FT \leq G \quad \textit{if and only if} \quad F \leq F_{\max}. \tag{14}$$

Moreover a necessary and sufficient condition for $F_{\max}$ to be non-singular is that T vanishes where G vanishes, i.e., that $\operatorname{Ker} G$ *is contained in* $\operatorname{Ker} T$.

Proof: See Section IV-B. □

Definition 12 The unique $F_{\max}$ with the properties of Theorem 11 is a positive-semidefinite $V \to V^*$ which we refer to as *the compression of G via T* denote by $T\backslash\backslash G := F_{\max}$.

The rest of this section is divided into two subsections. In subsection IV-A we state the Albert Condition for positivity [1] in the context of inner-products. This condition is used in subsection IV-B for the proof of Theorem 11.

A. The Albert Condition for Positivity

We shall first state an elementary fact which states that the inversion of inner-products is order reversing. A natural consequence of this, together with Theorem 23 of Section V-A, is that an increase of information results in a decrease of variance.

Lemma 13 *Let U denote a linear space equipped with two inner-products G and H. Then for all $\ell \in U^*$ there holds*

$$H_Q^{-*}(\ell) = \sup_{u\in U}\left\{\frac{|\ell(u)|^2}{H_Q(u)} : G_Q(u) = 1\right\}. \tag{15}$$

In particular $H \leq G$ is a necessary and sufficient condition for the inequality $G^{-*} \leq H^{-*}$.

Proof: To see that (7) and (15) are equivalent note that the normalization function $\nu : u \mapsto u/\sqrt{H_Q(u)}$ defines a bijection

$$\nu : \{u \,|\, G_Q(u) = 1\} \to \{u \,|\, H_Q(u) = 1\},$$

i.e., a correspondence between the unit circle of G_Q and the unit circle of H_Q. The "sufficiency"-part is clear from (15). To check necessity assume that $H \not\leq G$, i.e., there exists $v \in U$ such that $G_Q(v) < H_Q(v)$. Without loss of generality we may assume that $G_Q(v) = 1$ and apply (15) to obtain

$$\begin{aligned} H_Q^{-*}(Gv) &= \sup_{u\in U}\left\{\frac{|(Gv)(u)|^2}{H_Q(u)} : G_Q(u) = 1\right\} \\ &= \frac{G_Q(v)^2}{H_Q(v)} < G_Q(v) = G_Q^{-*}(Gv) \end{aligned}$$

where the last equality follows by Remark 5. This proves that $G^{-*} \not\leq H^{-*}$ and hence the "necessity"-part of the claim. □

In Theorem 14 we formulate Albert's positivity condition [1], [6] in the context of inner-products.

Theorem 14 *Let U and V denote two linear spaces equipped with inner-products given by*

$$G : U \to U^* \quad \textit{and} \quad F : V \to V^*,$$

respectively. For any surjective linear map $T : U \to V$ the following statements are equivalent:

1) $T^*FT \leq G$.
2) $T^{**}G^{-*}T^* \leq F^{-*}$.
3) $F \leq (TG^{-1}T^*)^{-1}$.

Proof: We first check that 1) is sufficient for 2) to hold. For this let $\ell \in V^*$. By Theorem 4 we can conclude that 2) holds via

$$\begin{aligned} F_Q^{-*}(\ell) &= \sup_{v\in V}\{|\ell(v)|^2 : F_Q(v) = 1\} \\ &= \sup_{u\in U}\{|\ell(Tu)|^2 : F_Q(Tu) = 1\} \\ &= \sup_{u\in U}\left\{\frac{|(T^*\ell)(u)|^2}{(T^*FT)_Q(u)} : G_Q(u) = 1\right\} \\ &\geq G_Q^{-*}(T^*\ell) = (T^{**}G^{-*}T^*)_Q(\ell). \end{aligned}$$

It remains to check that 1) is also necessary for 2) to hold. For this assume that 1) does not hold, i.e., $G_Q(u_0) < F_Q(Tu_0)$ for some $u_0 \in U$. Let $\ell_0 \in V^*$ be the unique solution of $T^*\ell_0 = Gu_0$. Without loss of generality we may assume that $G_Q(u_0) = 1$. By (15) and Remark 5 it follows that, indeed

$$\begin{aligned} F_Q^{-*}(\ell_0) &= \sup_{u\in U}\left\{\frac{|\ell(Tu)|^2}{F_Q(Tu)} : G_Q(u) = 1\right\} \\ &= \frac{|(Gu_0)(u_0)|^2}{F_Q(Tu_0)} \\ &< G_Q(u_0) = G_Q^{-*}(T^*\ell_0), \end{aligned}$$

i.e., $T^{**}G^{-*}T^* \not\leq F^{-*}$. It remains to check that 2) and 3) are equivalent. Since T is surjective, its adjoint T^* is injective. Hence $TG^{-1}T^*$ is indeed nonsingular. The equivalence of 2) and 3) then follows directly by applying Lemma 13. □

Remark 15 If T, the linear map defined in Theorem 14, is not surjective the condition $T^*FT \leq G$ remains sufficient for $T^{**}G^{-*}T^* \leq F^{-*}$. However for this condition to become necessary T must be surjective.

B. Existence and Uniqueness of the Compression

What now follows is a *Proof of Theorem 11*. It is worthwhile noting, that the proof is based on Albert's positivity condition for inner-products, i.e., on Theorem 14.

Proof: Let $N = \operatorname{Ker} G$ and $\bar{U} = U/N := \{[u] \,|\, u \in U\}$ denote the quotient space where the residue class $[u] \subseteq U$ of $u \in U$ modulo N is given by

$$[u] := u + N := \{u + n \,|\, n \in N\}, \tag{16}$$

Moreover let $P : U \to \bar{U}, u \mapsto [u]$ denote the natural projection. Since G is self-adjoint there exists a unique self-adjoint $\bar{G} : \bar{U} \to \bar{U}^*$ such that $G = P^*\bar{G}P$. Moreover by construction $\bar{G} \geq 0$ is non-singular. Assume first that T vanishes on N. Then $\bar{T} : \bar{U} \to V$ given by $\bar{T}[u] = Tu$ is well defined. Moreover by construction

$$\bar{T}^*F\bar{T} \leq \bar{G} \quad \Leftrightarrow \quad T^*FT \leq G. \tag{17}$$

The equivalence of statement 1) and 3) in Theorem 14 applied to the left hand side of (17) shows that the condition

$$F \leq F_{\max} \quad \text{with} \quad F_{\max} := (\bar{T}\bar{G}^{-1}\bar{T}^*)^{-1}, \tag{18}$$

is necessary and sufficient for $TFT^* \leq G$.

We now drop the assumption that T vanishes where G vanishes, i.e., we assume $\operatorname{Ker} T \not\supseteq N$. A necessary condition for $T^*FT \leq G$ is that F vanishes on $TN \subseteq V$. For each such F there exists a unique $\bar{F} : \bar{V} \to \bar{V}^*$ such that

$$F = Q^*\bar{F}Q \quad \text{where} \quad Q : V \to \bar{V}, \tag{19}$$

denotes the natural projection onto $\bar{V} := V/TN$ which denotes the quotient space. By construction we have that

$$T^*FT \leq G \quad \Leftrightarrow \quad (QT)^*\bar{F}(QT) \leq G.$$

Since QT vanishes on N it follows, by the first part of the proof, that there exists a unique inner-product $\bar{F}_{\max}$ on $\bar{V}$ such that the following equivalence

$$(QT)^*\bar{F}(QT) \leq G \quad \Leftrightarrow \quad \bar{F} \leq \bar{F}_{\max}, \tag{20}$$

holds. Since for all $F_i : \bar{V} \to \bar{V}^*$ there holds that

$$\bar{F}_1 \leq \bar{F}_2 \quad \Leftrightarrow \quad Q^*\bar{F}_1Q \leq Q^*\bar{F}_2Q, \tag{21}$$

we can conclude that $F_{\max} = Q^*\bar{F}_{\max}Q$ is the unique positive-semi-definite operator on V such that $F \leq F_{\max}$ is equivalent to $T^*FT \leq G$. It is clear that

$$\operatorname{Ker} F_{\max} = \operatorname{Ker} Q = TN, \tag{22}$$

and hence the pseudo inner-product $F_{\max}$ is non-singular if and only if T vanishes on $N = \operatorname{Ker} G$. □

C. The Lifting Property

We conclude this section with a lifting Theorem which relates the compressed inner-product to the original one.

Theorem 16 (Lifting Theorem) *Let U and V denote linear spaces, $G : U \to U^*$ and $T : U \to V$, with $G \geq 0$, and $T : U \to V$ surjective. If T vanishes where G vanishes, then*

$$(T\backslash\backslash G)_Q^{-*}(\ell) = \sup\{(T^*\ell)(u) \,|\, G_Q(u) = 1\}. \tag{23}$$

holds for all $\ell \in V^*$.

Proof: In the proof of Theorem 11 we saw that $T\backslash\backslash G$ is given by $(\bar{T}\bar{G}^{-1}\bar{T}^*)^{-1}$, i.e.,

$$(T\backslash\backslash G)^{-*} = \bar{T}^{**}\bar{G}^{-*}\bar{T}^*.$$

This together with Theorem 4 implies that

$$(\bar{T}^{**}\bar{G}^{-*}\bar{T}^*)_Q(\ell) = \sup\{|(\bar{T}^*\ell)([u])|^2 \,:\, \bar{G}_Q([u]) = 1\}.$$

By Definition of the equivalence class $[u]$ we have

$$G_Q(u) = \bar{G}_Q([u]),\ |(T^*\ell)(u)|^2 = |(\bar{T}^*\ell)([u])|^2,$$

and thus (23) indeed holds. □

V. Model spaces with statistical structure

An estimation problem is a special type of modeling problem where one observes random samples and seeks to estimate a model which explain the samples and yields structural insight into the random mechanism generating them.

In order to be able to find an estimate via optimization the set of possible estimates is constrained to a set $\mathcal{P}$ called *model space*. The model structure is determined before the estimation is performed but can be refined as more samples become available. The random samples are assumed to define elements in a measurable space X called *sample space* which is linked with the model space via a map

$$\mathbb{E} : \mathcal{P} \to \mathcal{E}(X^\infty), \quad P \mapsto \mathbb{E}_P, \tag{24}$$

which assign to a model $P \in \mathcal{P}$ a probability measure $\mathbb{E}_P$ on the sequence space

$$X^\infty = \{(x_1, x_2, \dots) \,|\, \forall t \in \mathbb{N} : x_t \in X\}. \tag{25}$$

Definition 17 Given a model space $\mathcal{P}$, a pair $(X, \mathbb{E})$ which satisfies (24) is said to define a *statistical structure* on $\mathcal{P}$ with samples in X.

Remark 18 Let $\mathbb{P} \in \mathcal{E}(X^\infty)$ denote a probability measure on X^∞ and $\mathbf{E}$ the corresponding expectation operator. Then for all events $F \subseteq X^\infty$ the probability $\mathbb{P}(F)$ is given by $\mathbf{E}[1_F]$ where 1_F is the indicator function of F which equals 1 on F and 0 elsewhere. It is therefore possible to identify probability measures with the corresponding expectation operators which we shall do from now on. Note that the notation $\mathbb{E}_P$ already suggests that we think of $\mathbb{E}_P$ as an expectation operator.

Definition 19 A model space is called *regular* if it forms a finite dimensional manifold over the real numbers. We then

refer to $n = \dim(\mathcal{P})$ as the *model order* and denote by $T_P\mathcal{P}$ the *tangent space* of $\mathcal{P}$ at P. A pair (k, γ) with $k \in \mathbb{K}$ and a curve $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{P}$ such that

$$\Delta(\alpha) := k^* \cdot \frac{\mathrm{d}}{\mathrm{d}t}\alpha(\gamma(t))^*\big|_{t=0} \quad \forall\, \alpha \in C^\infty(\mathcal{P}, \mathbb{K}), \tag{26}$$

is said to *represent* the tangent-vector $\Delta \in T_P\mathcal{P}$ with

$$\Delta : C^\infty(\mathcal{P}, \mathbb{K}) \to \mathbb{K}. \tag{27}$$

By virtue of these definitions the tangent-space $T_P\mathcal{P}$ is an n-dimensional linear space over $\mathbb{K}$.

For all smooth functions $\alpha \in C^\infty(\mathcal{P}, \mathbb{K})$ we define the *conjugate differential of α at $P \in \mathcal{P}$*, denoted by $\bar{\mathrm{D}}\alpha(P)$, as

$$(\bar{\mathrm{D}}\alpha(P))(\Delta) := \Delta(\alpha) \quad \text{for all} \quad \Delta \in T_P\mathcal{P}. \tag{28}$$

These definitions ensure that $\bar{\mathrm{D}}\alpha(P)$ is a conjugate linear functional, i.e., $\bar{\mathrm{D}}\alpha(P) \in (T_P\mathcal{P})^*$.

Finally if $\rho : \mathcal{P} \to \mathcal{R}$ is a smooth map between two regular model spaces we define via

$$(\mathrm{D}\rho(P))(\Delta) : C^\infty(\mathcal{R}, \mathbb{K}) \to \mathbb{K}, \quad \beta \mapsto \Delta(\beta \circ \rho), \tag{29}$$

for all $\Delta \in T_P\mathcal{P}$ and call $\mathrm{D}\rho(P) : T_P\mathcal{P} \to T_{\rho(P)}\mathcal{R}$ the *differential of ρ at P*.

Definition 20 Given a model space $\mathcal{P}$ which consists of functions $P : \Omega \to \mathbb{K}$ we call it a *function model space*. To emphasize that $\mathcal{P}$ is a function model space on Ω we write $(\mathcal{P}, \mathrm{ev}|_\Omega)$ where

$$\mathrm{ev}_z : \mathcal{P} \to \mathbb{K} \quad \text{with} \quad \mathrm{ev}_z(P) := P(z), \tag{30}$$

denotes the *evaluation* at $z \in \Omega$. In a function model space $(\mathcal{P}, \mathrm{ev}|_\Omega)$, where $\mathcal{P}$ is a regular model space, and all evaluations are smooth, we define $\mathrm{ev}_{P,z} := \mathrm{D}\,\mathrm{ev}_z(P)$ for all $z \in \Omega$, i.e., $\mathrm{ev}_{P,z}$ denotes the differential of ev_z at P. We call $(\mathcal{P}, \mathrm{ev}|_\Omega)$ a *regular function model space* if $(T_P\mathcal{P}, \mathrm{ev}_{P|\Omega})$ is a linear function space for all $P \in \mathcal{P}$.

In Section V-A we state the generalized information inequality given a possibly singular Fisher-information matrix based on C.R. Rao's observation. This matrix endows the model space with a Riemannian-metric [9], [8]. To treat the case where the Fisher-information matrix is singular we shall use the compression introduced in Section IV.

In Section V-B we specialize from regular model spaces to regular function model spaces. We use E.H. Moore's notion of positive kernels [7] to define the autocovariance function of a random function in a way which generalizes the definition of the covariance matrix of a random vector. The connection between variability and reproducing kernels in the context of system identification was first noted in [15].

A. The Generalized Information Inequality

Definition 21 Let $\mathcal{P}$ denote a regular model space with statistical structure $(X, \mathbb{E})$. An *estimator for $\mathcal{P}$ given* $(X, \mathbb{E})$ is a sequence $\hat{P} := (\hat{P}_N)_{N\geq 1}$, where for all N:

$$\hat{P}_N : X^\infty \to \mathcal{P} \quad \text{only depends on } x_1, \ldots, x_N. \tag{31}$$

We refer to $\hat{P}_N$ as the *estimator $\hat{P}$ given sample size N*.

Definition 22 For $P \in \mathcal{P}$ the *Fisher-information metric (FIM) given sample size N* is an inner-product on the tangent space $T_P\mathcal{P}$ denoted by $G_N(P)$ and defined as

$$G_N(P)_Q(\Delta) := \mathbb{E}_P\left|k^* \cdot \frac{\mathrm{d}}{\mathrm{d}t}\log f_N(x, \gamma(t))\big|_{t=0}\right|^2, \tag{32}$$

for $\Delta \in T_P\mathcal{P}$ realized by $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{P}$ with $\gamma(0) = P$ and $k \in \mathbb{K}$, where $f_N(x)$ denotes the probability density function of $(x_1, \ldots, x_N) \in X^N$. Moreover, one defines the *asymptotic FIM per sample (AFIM)*, denoted by $G_\infty(P)$, as the limit of $N^{-1}G_N(P)$ as the sample size $N \to \infty$.

The classical information inequality becomes a special case of our main result in Theorem 23 if one specializes to $\mathcal{R} = \mathcal{P}$ and $\rho(P) = P$. In this special case the compression of the FIM as defined in (34) is simply the FIM itself.

Theorem 23 *Let $\hat{P} = (\hat{P}_N)_{N\geq 1}$ denote an estimator for a regular model space $\mathcal{P}$ given a statistical structure $(X, \mathbb{E})$. Moreover assume that $\mathcal{R}$ is a regular model space which denotes the image of a smooth map*

$$\rho : \mathcal{P} \to \mathcal{R}, \quad P \mapsto R, \tag{33}$$

with surjective derivative. Let $\beta : \mathcal{R} \to \mathbb{K}$ denote a smooth function and let $\alpha(P) := \beta(\rho(P))$, i.e., $\alpha = \beta \circ \rho$, such that $\mathcal{P} \to \mathbb{K}$. We denote the FIM given N samples and the asymptotic average FIM compressed by $\mathrm{D}\rho(P)$ as

$$G_{\rho\backslash\backslash N}(P) := \mathrm{D}\rho(P)\backslash\backslash G_N(P), \tag{34a}$$
$$G_{\rho\backslash\backslash \infty}(P) := \mathrm{D}\rho(P)\backslash\backslash G_\infty(P), \tag{34b}$$

respectively. For all N where the compressed FIM given N samples is positive definite, and $\alpha(\hat{P}_N)$ is unbiased, i.e.,

$$\mathbb{E}_P[\alpha(\hat{P}_N)] = \alpha(P) \quad \textit{for all} \quad P \in \mathcal{P}, \tag{35}$$

the information inequality is given by:

$$\mathbb{E}_P|\beta(\hat{R}_N) - \beta(R)|^2 \leq G_{\rho\backslash\backslash N}(P)^{-*}_Q(\bar{\mathrm{D}}\beta(R)), \tag{36}$$

where $\hat{R}_N = \rho(\hat{P}_N)$ and $R = \rho(P)$.

Proof: Let $f_N(x, P)$ denote the probability density function of $(x_1, \ldots, x_N) \in X^N$ when the law of x is $\mathbb{E}_P$ and $\hat{\alpha}_N := \alpha \circ \hat{P}_N$. Since $\hat{\alpha}_N$ is an unbiased estimator of α for any $\Delta \in T_P\mathcal{P}$ realized by $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{P}$, $k \in \mathbb{K}$ we have

$$(\bar{\mathrm{D}}\alpha(P))(\Delta) = \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}_{\gamma(t)}[k^*\hat{\alpha}_N]\big|_{t=0}, \tag{37}$$

which by "differentiation under the integral" becomes

$$\mathbb{E}_P\left[\hat{\alpha}_N(x) \cdot k^* \frac{\mathrm{d}}{\mathrm{d}t}\log f_N(x, \gamma(t))\big|_{t=0}\right]. \tag{38}$$

From this one deduces that $(\bar{\mathrm{D}}\alpha(P))(\Delta)$ equals

$$\mathbb{E}_P\left[(\hat{\alpha}_N(x) - \alpha(P)) \cdot k^* \frac{\mathrm{d}}{\mathrm{d}t}\log f_N(x, \gamma(t))\big|_{t=0}\right]. \tag{39}$$

It now follows from (32) and the Cauchy-Bunyakovsky-Schwarz-inequality that

$$|(\bar{\mathrm{D}}\alpha(P))(\Delta)|^2 \leq G_N(P)_Q(\Delta) \cdot \mathbb{E}_P|\hat{\alpha}_N - \alpha(P)|^2. \tag{40}$$

In other words the expression

$$\sup\{|(\bar{\mathrm{D}}\alpha(P))(\Delta)|^2 \,:\, G_N(P)_Q(\Delta) = 1\}, \qquad (41)$$

yields a lower bound for $\mathbb{E}_P|\hat{\alpha}_N - \alpha(P)|^2$. By the chain rule for differentiation we have

$$\bar{\mathrm{D}}\alpha(P) = \bar{\mathrm{D}}\beta(\rho(P)) \circ \mathrm{D}\rho(P), \qquad (42)$$

which by the lifting Theorem 16 implies (36). □

B. Information Inequality on Function Spaces

Definition 24 A *positive kernel* on a set Ω is a two variable function $\Phi : \Omega \times \Omega \to \mathbb{K}$, which is conjugate symmetric, i.e., satisfies $\Phi(z,w) = \Phi(w,z)^*$, such that the inequality

$$\sum\nolimits_{i,j=1}^{J} \Phi(z_i, z_j)\, \xi_j \xi_i^* \geq 0 \quad \text{where} \quad z_i \in \Omega, \xi_i \in \mathbb{K}, \quad (43)$$

is satisfied for all integers J. One writes $\Phi \geq 0$ to denote that Φ is a positive kernel. If Ψ is another positive kernel one writes $\Phi \leq \Psi$ or $\Psi \geq \Phi$ for $\Psi - \Phi \geq 0$.

Corollary 25 *Let $(\mathcal{R}, \mathrm{ev}|_\Omega)$ denote a regular function model space which satisfies the assumptions of Theorem 23. Define $\hat{R}_N := \rho \circ \hat{P}_N$ and $\mathrm{Cov}_P(\hat{R}_N) : \Omega \times \Omega \to \mathbb{K}$ via*

$$\mathrm{Cov}_P(\hat{R}_N)(z,w) := \mathbb{E}_P[\Delta(z)\Delta(w)^*], \qquad (44a)$$

$$\textit{where} \quad \Delta(z) := \mathrm{ev}_z(\hat{R}_N) - \mathbb{E}_P[\mathrm{ev}_z(\rho(P))], \qquad (44b)$$

for all $z, w \in \Omega$. Let N be such that

$$G_{\rho\backslash\backslash N}(P) > 0 \quad \textit{and} \quad \mathbb{E}_P[\mathrm{ev}_z(\hat{R}_N)] = \mathrm{ev}_z(R), \qquad (45)$$

with $R = \rho(P)$, i.e., the compressed FIM is positive definite and $\hat{R}_N$ is unbiased. The information inequality is given by

$$\mathrm{Cov}_P(\hat{R}_N) \geq K_{\rho\backslash\backslash N}(P), \qquad (46)$$

where $K_{\rho\backslash\backslash N}(P)$ is the reproducing kernel of $(T_R\mathcal{R}, \mathrm{ev}_{R|\Omega})$ with respect to $G_{\rho\backslash\backslash N}(P)$.

Proof: Let $J \in \mathbb{N}$, $\xi_1, \ldots, \xi_J \in \mathbb{K}$ and $z_1, \ldots, z_J \in \Omega$ be arbitrary. Moreover let

$$\beta : \mathcal{R} \to \mathbb{K} \quad \text{with} \quad \beta = \sum\nolimits_{j=1}^{J} \xi_j \cdot \mathrm{ev}_{z_j}.$$

Then there holds that

$$\sum\nolimits_{i,j=1}^{J} \mathrm{Cov}(\hat{R}_N)(z_i, z_j)\xi_j\xi_i^* = \mathbb{E}_P|\beta(\hat{R}_N) - \beta(R)|^2.$$

By Theorem 23 there holds that

$$\mathbb{E}_P|\beta(\hat{R}_N) - \beta(R)|^2 \leq G_{\rho\backslash\backslash N}(P)^{-*}_Q(\bar{\mathrm{D}}\beta(R)).$$

By definition $\mathrm{ev}_{R,z} = \overline{\bar{\mathrm{D}}\,\mathrm{ev}_z(P)}$ for all $z \in \Omega$ which implies

$$\bar{\mathrm{D}}\beta(R) = \sum\nolimits_{j=1}^{J} \xi_j^* \cdot \overline{\mathrm{ev}}_{R,z_j},$$

and thus $G_{\rho\backslash\backslash N}(P)^{-*}_Q(\bar{\mathrm{D}}\beta(R))$ equals

$$\sum\nolimits_{i,j=1}^{J} G_{\rho\backslash\backslash N}(P)^{-*}_Q(\overline{\mathrm{ev}}_{R,z_i}, \overline{\mathrm{ev}}_{R,z_j})\xi_j\xi_i^*.$$

This proves that indeed $\mathrm{Cov}_P(\hat{R}_N) \geq K_{\rho\backslash\backslash N}(P)$ in the sense of E.H. Moore. □

Corollary 25 gives a list of the ingredients which shape the autocovariance function of a reduced order estimator. The information metric $G_N(P)$ influences the autocovariance function modulo the kernel of the differential of the model reduction map. The precise influence is captured in the compressed information $G_{\rho\backslash\backslash N}$. The model structure $\mathcal{R}$ influences the autocovarinace function by the shape of its tangent-space. Given a desired autocovariance function it is therefore of interest to find a model-reduction map onto a reduced order model structure with a suitable tangent-space. In order to render this strategy sucessful one must avoid adding too much bias. For this to be possible one has to collect a priori information about the true system.

VI. 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 asymptotically efficient. The consequence of these results 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 Fisher-information metric. Using the novel concept of compression for inner-products we were able to treat the general case where the system need is not necessarily an element in the model set and the Fisher-information matrix is allowed to be singular.

This work can be easily extended to the multiple-input multiple-output case by replacing the scalar valued evaluation functionals by vector valued ones. In the future we plan to exploit the connection to geometric results, like [10], which give concrete functional models for tangent spaces of transfer function manifolds for SISO and MIMO systems.

VII. Acknowledgments

The authors would like to thank Brian D.O. Anderson and Paul Van Dooren for helpful discussions.

References

[1] A. Albert. Conditions for positive and nonnegative definiteness in terms of pseudoinverses. *SIAM Journal on Applied Mathematics*, 17(2):434–440, March 1969.

[2] H. Hjalmarsson, M. Gevers, and F. de Bruyne. For model-based control design, closed-loop identification gives better performance. *Automatica*, 32(12):1659–1673, 1996.

[3] J. Anderson, W. N. and G. E. Trapp. Shorted operators. ii. *SIAM Journal on Applied Mathematics*, 28(1):60–71, January 1975.

[4] P. Jordan and J. V. Neumann. On inner products in linear, metric spaces. *Annals of Mathematics*, 36(3):719–723, July 1935.

[5] J. M. Oller and J. M. Corcuera. Intrinsic analysis of statistical estimation. *The Annals of Statistics*, 23(5):1562–1581, October 1995.

[6] F. Zhang, editor. *The Schur Complement and Its Applications*, volume 4 of *Numerical Methods and Algorithms*. Springer-Verlag, New York, 2005.

[7] E. H. Moore. On properly positive Hermitian matrices. *Bull. Amer. Math. Soc.*, 23(59):66–67, 1916.

[8] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry (Translations of Mathematical Monographs)*, volume 191. American Mathematical Society, April 2001.

[9] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 37:81–91, 1945.

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

[11] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, 1998.

[12] A. A. Borovkov. *Mathematical Statistics*. Gordon and Breach Science Publish, 1998.

[13] J. M. Oller and J. M. Corcuera. Intrinsic analysis of statistical estimation. *The Annals of Statistics*, 23(5):1562–1581, October 1995.

[14] C. Carmeli, E. D. Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4:377–408, 2006.

[15] B. Ninness and H. Hjalmarsson. Variance error quantifications that are exact for finite-model order. In *Transactions on Automatic Control*, volume 49, pages 1275–1290, August 2004.