Information Inequality for Estimation of Transfer Functions: main results *

Tzvetan Ivanov ∗ Brian D.O. Anderson ∗∗ P.-A. Absil ∗ Michel Gevers ∗

∗ Center for Systems Engineering and Applied Mechanics (CESAME)
Universite Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium.
(tzvetan.ivanov@uclouvain.be, michiel.gevers@uclouvain.be,
absil@inma.ucl.ac.be)

** National ICT Australia and the Australian National University
Research School of Information Sciences & Engineering, Canberra,
ACT 2601, Australia (brian.anderson@anu.edu.au)

Abstract: In this paper we derive a canonical lower bound for the autocovariance function of any unbiased transfer-function estimator. As a generalization of the Cramér-Rao bound, the Cramér-Rao kernel that we define can be derived without parametrizing the model set. The Cramér-Rao kernel is thus one of the cornerstones for experiment design formulations that do not depend on the choice of coordinates.

Keywords: System Identification, Metric, Confidence Region, Autocovariance, Reproducing Kernel, Fisher Information, Information Geometry

1. INTRODUCTION

Parameter estimation enjoys a rich set of features which makes it attractive in the context of system identification (SYSID). One of the main features is that parameter estimation, when used in a probabilistic framework, allows users to quantify their confidence in how closely the estimated parameter matches the true one. For example in experiment design one uses classical results, such as the Cramér-Rao lower bound (CRLB), in order to tune the experimental conditions in such a way that the achieved confidence region matches some performance specification. However, finding the optimal experimental conditions can be challenging if the performance specification is not directly given in the parameter space. For example, if the goal is to estimate a transfer function, then the performance specification is not directly given in the parameter space. However, it is typically given in terms of a maximal tolerable distance between the true transfer function and the estimated transfer function and not in terms of a maximal tolerable distance between the corresponding parameter vectors. In order to execute the experiment design one can often use the fact that they are related via a bijective function – the parametrization.

Additionally, the notion of distance between the true transfer function and the estimated transfer function depends on the application, and is often induced by norms on the embedding space such as the $L_∞$-norm or $L_2$-norm. In general, and perhaps regretfully, natural distance functions on the model set lead to complicated, non-linear, distance functions on the parameter space.

We are therefore interested to circumvent the usage of parametrizations and replace the main players in the optimal experiment design problem. To be concrete let $C(\hat{\theta}) \in \mathbb{R}^{d \times d}$ denote the covariance matrix of an unbiased estimator $\hat{\theta}$ of a parameter $\theta^* \in \Theta \subseteq \mathbb{R}^d$, and assume the Fisher-information matrix, call it $J(\theta^*) \in \mathbb{R}^{d \times d}$, is a non-singular matrix. Then the problem is how to replace the CRLB, i.e., the matrix inequality

$$C(\hat{\theta}) \geq J(\theta^*)^{-1}, \quad (1)$$

by an inequality between two objects which are defined purely in terms of the model set and the probabilistic assumption, i.e., in a manner which does not involve parametrizations at all. In most cases then, we are seeking a replacement of (1) by an inequality somehow reflecting transfer functions.

One part of this problem has been solved in (Rao, 1945), where one showed that the information matrix $J(\theta)$ at a fixed parameter $\theta^* \in \Theta$ satisfies

$$[J(\theta^*)]_{ij} = \left( \frac{\partial \Pi(\theta^*)}{\partial \theta^*_j}, \frac{\partial \Pi(\theta^*)}{\partial \theta^*_i} \right)_{P_*}, \quad (2)$$

where $(\cdot, \cdot)_{P_*}$ denotes an inner-product at the tangent space of the model set $\mathcal{P}$ at the unique model $P_* \in \mathcal{P}$ such that $\Pi = \Pi(\theta^*)$ holds for the parametrization $\Pi : \Theta \to \mathcal{P}$. The key point is that this inner-product is actually independent of the chosen parametrization and, when viewed as a function of $P_*$, defines a metric, the so called information metric, on the model set, cf. (Amari
and Nagaoka, 2001). This observation alone, however, does not suffice to solve the original problem which requires parametrization independent definitions for the covariance matrix of the estimated quantity. The inversion of the information matrix and also the standard matrix ordering used in (1) has to be redefined in an appropriate way.

The contribution of this paper is the derivation of a coordinate-free version of the CRLB (1) for the case where:

- The elements \( P \in \mathcal{P} \) of the model set \( \mathcal{P} \) are complex-valued functions \( P : \mathbb{T} \rightarrow \mathbb{C} \), defined on the unit circle \( \mathbb{T} \) of the complex plane \( \mathbb{C} \), i.e., discrete-time transfer functions.
- The model set \( \mathcal{P} \) is a \( d \)-dimensional differentiable manifold over the real numbers \( \mathbb{R} \), i.e., the set of transfer functions admits a smooth parametrization defined on a subset of \( \mathbb{R}^d \).
- The information metric \( \langle \cdot, \cdot \rangle_{\mathcal{P}} \) is non-singular, i.e., the data set is sufficiently rich that unique identification can be achieved.

We show that for every function estimator \( \hat{P} \) such that

\[
\mathbb{E}_{\mathcal{P}}[\hat{\eta}_P(z)] = P(z) \quad \text{for all} \quad z \in \mathbb{T}, P \in \mathcal{P},
\]

(i.e., \( \hat{P} \) is unbiased), the corresponding autocovariance function admits a lower bound designated \( K_{\mathcal{P}} \), below.

The autocovariance function is the two variable function defined for all \( z, w \in \mathbb{T} \) via

\[
\text{Cov}_{\mathcal{P}}(\hat{P})(z, w) := \mathbb{E}[(\hat{P}(z) - P(z)) \cdot (\hat{P}(w) - P(w))].
\]

By lower bound we mean an inequality of the form

\[
\sum_{k,l=1}^{M} \eta_k \eta_l \cdot \left( \text{Cov}_{\mathcal{P}}(\hat{P}(z_k, z_l) - K_{\mathcal{P}}(z_k, z_l)) \right) \geq 0,
\]

which holds for every integer \( M \geq 0 \), every set of evaluation points \( z_1, \ldots, z_M \in \mathbb{T} \), and every vector \( \eta \in \mathbb{C}^M \). The inequality (5) simply means

\[
\text{Cov}_{\mathcal{P}}(\hat{P}) \geq_{EH} K_{\mathcal{P}},
\]

where \( \geq_{EH} \) denotes the E.H. Moore ordering (Moore, 1916).

We show that the classical CRLB corresponds to the case where \( K_{\mathcal{P}} \) is the reproducing kernel (cf. Appendix A) of the tangent-space of the model set at \( P \), with respect to the information metric \( \langle \cdot, \cdot \rangle_{\mathcal{P}} \) defined in (2). Due to its fundamental importance in variance quantification we call \( K_{\mathcal{P}} \), the Cramér-Rao kernel.

The paper is structured as follows. In Section 2 and 3 we develop the Cramér-Rao kernel lower bound starting from a parametric setup. In Section 4 we prove reparametrization invariance of the previous results which will allow us to replace them, in Section 5, by purely geometric definitions which do not involve parametrizations. In Section 6 we demonstrate the effectiveness of our approach by revisiting asymptotic sample size analysis in system identification for the special model set \( \mathcal{P} = \text{Rat}(n) \) which consists of all transfer functions with McMillan degree \( n \). In Section 7 we give a sample application arising from performance specifications used in robust control. In Section 8 we conclude.

Notation:
- \( \mathbb{R} \) real numbers, \( \mathbb{C} \) complex numbers, \( z \) complex conjugate, \( j = \sqrt{-1} \) imaginary unit, \( A^T \) transpose of a matrix \( A \), \( A^H \) conjugate-transpose of a matrix \( A \), \( \mathbb{T} \) unit-circle, \( \mathbb{E} \) expectation

2. THE CRAMÉR-RAO KERNEL

In this section we shall derive the Cramér-Rao kernel which plays the role of the inverse information matrix in the context of function estimation. To highlight the connections with the classical parametric approach in Lemma 2.1 we derive the Cramér-Rao kernel given a fixed parametrization \( \Pi \) of the function space \( \mathcal{P} \).

Before we continue, we first introduce the setup and fix the notation for the rest of the paper:

- parameter-space \( \Theta \subset \mathbb{R}^d \), parameter-vector \( \theta \in \Theta \),
- sample-space \( X \subset \mathbb{R}^d \), random sample \( 1 \times X \),
- parametrized set of probability density functions \( \mathcal{D} = \{ p_\theta : X \rightarrow \mathbb{R}_{\geq0} \mid \theta \in \Theta \} \),
- expectation \( \mathbb{E}_\theta[f(x)] \) of a particular measurement \( f : X \rightarrow \mathbb{C} \) w.r.t. to a random sample \( x \sim p \) with density \( p \in \mathcal{D} \),
- parametrized model set of transfer functions
  \[ \mathcal{P} = \{ P : \mathbb{T} \rightarrow \mathbb{C} \mid P = \Pi(\theta), \theta \in \Theta \}, \]
- \( z \in \mathbb{T} \) evaluation point, evaluated parametrization \( \Pi_z : \Theta \rightarrow \mathbb{C} \) with \( \Pi_z(\theta) := (\Pi(\theta))(z) \),
- parameter estimator \( \hat{\theta}(x) \in \Theta \), function estimator \( \hat{P}_z \in \mathcal{P} \) with \( \hat{P}_z(z) = \Pi_z(\hat{\theta}(x)) \),
- unbiasedness for function estimators \( \mathbb{E}_{\Pi_\theta}[\hat{P}_z(z)] = \Pi_z(\theta) \) for all \( z \in \mathbb{T}, \theta \in \Theta \),
- autocovariance of an unbiased estimator \( \hat{P}_z \)
  \[ \text{Cov}_{\Pi_\theta}(\hat{P}_z)(z, w) := \mathbb{E}_{\Pi_\theta}[(\hat{P}_z(z) - \Pi_z(\theta)) \cdot (\hat{P}_z(w) - \Pi_w(\theta))], \]
  for all \( z, w \in \mathbb{T} \) and \( \theta \in \Theta \).

We shall make the following regularity assumptions:

R1) The model set \( \mathcal{P} \) is a differentiable manifold, the parameter space \( \Theta \) is open, and \( \Pi \) is a smooth parametrization of \( \mathcal{P} \).

R2) For every fixed \( z \in X \) the function \( \Theta \rightarrow \mathbb{R}_{\geq0}, \theta \mapsto p_\theta(z) \) is smooth. Similarly for every fixed \( z \in \mathbb{T} \) the function \( \Theta \rightarrow \mathbb{C}, \theta \mapsto \Pi_z(\theta) \) is well-defined and smooth.

R3) For every \( \theta^* \in \Theta \) the information matrix \( J_\theta \) given by
  \[ [J_\theta]_{ij} = \mathbb{E}_{\Pi_{\theta^*}} \left[ \frac{\partial \log p_{\theta^*}(x)}{\partial \theta_j} \cdot \frac{\partial \log p_{\theta^*}(x)}{\partial \theta_i} \right], \]
  \( i, j = 1, \ldots, d \), is non-singular.

R1 states that the parameter vector \( \theta \) is globally-identifiable, i.e., that \( \Pi(\theta) = \Pi(\theta^*) \) implies that \( \theta = \theta^* \) (Gevers et al., 2009). R3 ensures that the samples are sufficiently rich to distinguish between parameters.

1 A random sample is like a sequence of measurements on the system, for example a sequence of input and output values \( y_k, x_k, k = 1, 2, \ldots, N \) for a linear plant with outputs contaminated by additive noise, defined by an equation like \( y = Pu + v \) the symbols having obvious meaning.

2 In particular we require that for all \( \theta \in \Theta \) the function \( P = \Pi(\theta) \) is defined for all \( z \in \mathbb{T} \). If \( P \) is a rational function this is equivalent to saying that \( P \) has no pole on the unit circle.
Remark 2.1. Due to assumption R1), we may abuse notation and write $E_P$ and $Cov_P$ instead of $E_p$ and $Cov_p$ whenever there holds $p = p_0$ and $P = \Pi(\theta)$.

Lemma 2.1. Let $\hat{P}_x \in \mathcal{P}$ denote an unbiased estimator in the sense of (11); then (5) holds with
\[
K_{\mathcal{P}}(z, w) = \sum_{i,j=1}^{d} [J_{\theta}^{-1}]_{ji} \cdot \frac{\partial \Pi_i(\theta^*)}{\partial \theta_i} \cdot \frac{\partial \Pi_j(\theta^*)}{\partial \theta_j},
\]
where $P_\alpha = \Pi(\theta^*)$. We call $K_{\mathcal{P}}$ the $J_\theta$-induced Cramér-Rao kernel at the model $P_\alpha$.

Proof. We first note that
\[
\sum_{k,l=1}^{M} \eta_k \Pi_{\alpha} \Pi_{\alpha} \cdot \frac{\partial \Pi_k(\theta)}{\partial \theta_l} \cdot \frac{\partial \Pi_l(\theta)}{\partial \theta_k} = \mathbb{E}_{\mathcal{P}}[\hat{\alpha}(x) - \alpha(\theta)]^2,
\]
which, together with equation (15), is equivalent to (5). □

3. THE REPRODUCING PROPERTY

In the previous section we saw that the $J_\theta$-induced Cramér-Rao kernel, i.e., the two variable function $K_{\mathcal{P}}(z, w)$ defined in (14) yields a lower bound on the auto-covariance of any unbiased function estimator $\hat{P}$ of $P_\alpha = \Pi(\theta^*)$.

In this section we will introduce the information metric and study its relation to the Cramér-Rao kernel. The main result of this section is that the Cramér-Rao kernel has the reproducing property (cf. Appendix A) with respect to the information metric.

In order to highlight the geometric aspects which are inherent in our setup, especially to the fact that the model set $\mathcal{P}$ is a differentiable manifold, we shall now adapt the standard information geometry terminology to suit our needs for function spaces.

Definition 3.1. For $\theta^* \in \Theta$ we define the tangent space of $\mathcal{P}$ at $P_\alpha = \Pi(\theta^*)$ via
\[
T_{\mathcal{P}} \mathcal{P} := \{ \Delta : \mathbb{T} \to \mathbb{C} \mid \exists \delta \in \mathbb{R}^d \text{ such that } \forall z \in \mathbb{T} : \Delta(z) = \sum_{i=1}^{d} \delta_i \frac{\partial \Pi_i(\theta^*)}{\partial \theta_i} \}.
\]

The $J_\theta$-induced information metric on $\mathcal{P}$ at $P_\alpha$ is the inner-product $\langle , \rangle : T_{\mathcal{P}} \mathcal{P} \times T_{\mathcal{P}} \mathcal{P} \to \mathbb{R}$ defined by
\[
\left(\Delta^{(1)}, \Delta^{(2)}\right)_{P_\alpha} := \delta^{(2),T} J_\theta \delta^{(1)},
\]
for all $\Delta^{(1)}(z) = \sum_{i=1}^{d} \delta_i \frac{\partial \Pi_i(\theta^*)}{\partial \theta_i}$ with $J_\theta$ as in (13).

In Theorem 3.1 we shall prove the reproducing property of the $J_\theta$-induced Cramér-Rao kernel with respect to the $J_\theta$-induced information metric.

Theorem 3.1. Let $w \in \mathbb{T}$ be fixed. Then there exist uniquely defined functions $U, V \in T_{\mathcal{P}} \mathcal{P}$ in the tangent space such that for all $\Delta \in T_{\mathcal{P}} \mathcal{P}$ there holds
\[
\Delta(w) = \left(\Delta, U\right)_{P_\alpha} - j \cdot \left(\Delta, V\right)_{P_\alpha}.
\]
The Cramér-Rao kernel is then uniquely determined by
\[
K_{\mathcal{P}}(z, w) = U(z) + j \cdot V(z) \quad \text{for all } z \in \mathbb{T}.
\]
We call this the reproducing property of $K_{\mathcal{P}}$ with respect to the $J_\theta$-induced information metric.

Proof. The fact that $U,V$ are uniquely defined follows from the Riesz-representation theorem for Hilbert spaces. Let $\Delta(z) = \langle \Delta, w \rangle_{P_\alpha}$ for all $z \in \mathbb{T}$ and $i = 1, \ldots, d$. To check (19) we note that
\[
K_{\mathcal{P}}(z, w) = \sum_{i,j=1}^{d} [J_{\theta}^{-1}]_{ji} \Delta_i(z) \cdot \overline{\Delta_j(w)},
\]
and thus
\[
K_{\mathcal{P}}(z, w) = R(z) + J(z)
\]
with
\[
\left\{ R(z) = \sum_{i,j=1}^{d} [J_{\theta}^{-1}]_{ji} \Delta_i(z) \cdot \text{Re}(\overline{\Delta_j(w)}),
\right.
\left. J(z) = \sum_{i,j=1}^{d} [J_{\theta}^{-1}]_{ji} \Delta_i(z) \cdot \text{Im}(\overline{\Delta_j(w)}),
\right\}
\]
for all $z \in \mathbb{T}$.

Let $\Delta \in T_{\mathcal{P}} \mathcal{P}$ denote a tangent-vector. We expand $\Delta = \sum_{i=1}^{d} \delta_i \cdot \Delta_i$ with $\delta \in \mathbb{R}^d$, and compute
\[
\Delta(w) = \left[\Delta_1(w), \ldots, \Delta_d(w)\right] \delta = \left[\text{Re} \Delta_1(w), \ldots, \text{Re} \Delta_d(w)\right] \delta + j \cdot \left[\text{Im} \Delta_1(w), \ldots, \text{Im} \Delta_d(w)\right] \delta
\]
and
\[
\Delta(w) = \left(\Delta, U\right)_{P_\alpha} - j \cdot \left(\Delta, V\right)_{P_\alpha}.
\]
By the uniqueness of $U,V$ it follows that $R = U$ and $J = V$. Together with equation (20) this implies that indeed (19) holds. □

4. REPARAMETRIZATION INVARIANCE

In this section we shall show that the lower bound $K_{\mathcal{P}}$ defined in (14), which we obtained in Lemma 2.1, is independent of the chosen parametrization $\Pi : \Theta \to \mathcal{P}$.

For this let $\Xi \subseteq \mathbb{R}^d$ denote an open subset such that
\[
\mathcal{D} = \{ \hat{p}_\xi \mid \xi \in \Xi\} \quad \text{and} \quad \mathcal{P} = \{ \Pi(\xi) \mid \xi \in \Xi\},
\]
denote alternative parametrizations of the set of densities and model set defined by (7) and (8), respectively, which obey the regularity assumptions R1) to R3). By assumption R1) there exists a unique smooth map $\varphi : \Xi \to \Theta$, the chart-transition, such that
\[
\Pi(\xi) = \Pi(\varphi(\xi)) \quad \text{for all } \xi \in \Xi.
\]

In order to proceed we have to make a probabilistic compatibility assumption
\[
\hat{p}_\xi = p_{\varphi(\xi)} \quad \text{for all } \xi \in \Xi.
\]
such that the abuse of notation in Remark 2.1 remains admissible.
Lemma 4.1. For \( \xi^* \in \Xi \) and \( P^* = \tilde{\Pi}(\xi^*) \) define
\[
[\tilde{J}_{\xi^*}]_{ij} = E_{\tilde{p}_{\xi^*}} \left[ \frac{\partial \log \tilde{p}_{\xi^*}(x)}{\partial \xi_i} \cdot \frac{\partial \log \tilde{p}_{\xi^*}(x)}{\partial \xi_j} \right].
\] (23)
Let \( \theta^* \in \Theta \) such \( \Pi(\theta^*) = P^* \), then
- There holds that
  \[
  K_{P^*}(z, w) = \sum_{i,j=1}^d \left[ J^{(1)}_{\xi^*} \right]_{ij} \cdot \delta^2 \Pi_{\xi^*}(z) \cdot \delta^2 \Pi_{\xi^*}(w),
  \] (24)
i.e., the \( J_{\xi^*} \)-induced Cramér-Rao kernel (the right hand side of (24)) and \( J_{\theta^*} \)-induced Cramér-Rao kernel defined in (14) are equal.
- The tangent-space (16) can be spanned using \( \Pi \) instead of \( \Pi \), i.e.,
  \[
  T_{P^*} \mathcal{P} = \{ \Delta : \mathbb{T} \to \mathbb{C} | \exists \delta \in \mathbb{R}^d \}
  \] such that
  \[
  \forall z \in \mathcal{T} : \Delta(z) = \sum_{i=1}^d \delta^i \Pi_{\xi^*}(z),
  \]
i.e., the \( \Delta^1 \)-induced information metric (17) and the \( J_{\xi^*} \)-induced information metric (23) are equal.

Proof. The proof follows by a change of variables and the chain rule of differentiation. The details are omitted.

5. COORDINATE-FREE DEFINITIONS

We recall that in Section 2 and 3 we defined the Cramér-Rao kernel \( K_{P^*} \) in (14), the tangent-space \( T_{P^*} \mathcal{P} \) in (16), and the information metric \( (\cdot, \cdot)_{P^*} \) in (17) w.r.t. a fixed parametrization. In Section 4 we proved that the resulting definitions are invariant if one uses another parametrization \( \tilde{\Pi} : \Xi \to \mathcal{P} \) instead of \( \Pi \). Our goal in this section will be to define the Cramér-Rao kernel and the information metric without invoking a parametrization.

There are various ways to introduce the notion of tangent space in a coordinate-free way (Lang, 2002). We choose the following: For \( \mathcal{P} \subset \mathbb{P} \) the tangent space of \( \mathcal{P} \) at \( P \) is
\[
T_{P} \mathcal{P} := \{ \Delta : \mathbb{T} \to \mathbb{C} | \exists \delta \in \mathbb{R}^d \}
\]
where \( \forall z \in \mathbb{T} : \Delta(z) = \sum_{i=1}^d \delta^i \Pi_{\xi^*}(z) \)
and the curve \( \gamma \) is always assumed smooth. This definition is equivalent to the one given by (16).

Definition 5.1. Let \( \mu : \mathcal{P} \to \mathcal{D}, \mu_{P^*} \) denote a fixed smooth map. With respect to \( \mu \) one defines the information metric via
\[
E_{\mu_{P^*}} \left[ \frac{d}{d\tau} \log \mu_{\gamma}(\tau) \right]_{\tau=0} := (\Delta_{\gamma_1}, \Delta_{\gamma_2})_{P^*} := \left( \frac{d}{d\tau} \log \mu_{\gamma}(\tau) \right)_{\tau=0},
\] (27)
for all \( \Delta_{\gamma} \in T_{P^*} \mathcal{P} \). We assume that the information metric is non-singular.\(^3\)

For each fixed \( w \in \mathbb{T} \) there exist uniquely defined functions \( U_w, V_w \in T_{P^*} \mathcal{P} \) such that
\[
\begin{align*}
\text{Re} \Delta(w) &= (\Delta, U_w)_{P^*}, \\
\text{Im} \Delta(w) &= -(\Delta, V_w)_{P^*},
\end{align*}
\] (28)
for all \( \Delta \in T_{P^*} \mathcal{P} \).

\(^3\) The metric \( (\cdot, \cdot)_{P^*} \) is called non-singular if \( (\Delta, \Delta)_{P^*} = 0 \) implies that \( \Delta = 0 \) for all \( \Delta \in T_{P^*} \mathcal{P} \).

The function defined by
\[
K_{P^*}(z, w) := U_w(z) + jV_w(z)
\] for all \( z, w \in \mathbb{T} \), (29) is called the Cramér-Rao kernel.

In Theorem 5.1 we establish that the nomenclature in Definition 5.1 is consistent with the earlier definitions made in Section 2 and Section 3.

Theorem 5.1. Assume that \( \mu : \mathcal{P} \to \mathcal{D} \) is such that the probabilistic compatibility assumption
\[
\mu_{\Pi(\theta)} = p_{\theta}
\] holds for all \( \theta \in \Theta \).

Proof. Let \( \epsilon_i = (0, \ldots, 0), \ i = 0 \), \( 0, \ldots, 0 \) \( \in \mathbb{R}^d \) and
\[
\gamma_i : (-1, 1) \to \mathcal{P},
\] with \( \gamma_i(\tau) = \Pi(\theta, \tau \cdot \epsilon_i) \).

Then one computes
\[
[J_{\gamma}]_{ij} = E_{\mu_{P^*}} \left[ \frac{d}{d\tau} \log \mu_{\gamma_i}(\tau) \right]_{\tau=0}
\]
using the compatibility assumption (18). The equivalence of (14) and (29) then follows from the reproducing property proved in Theorem 3.1.

6. APPLICATION TO SYSTEM IDENTIFICATION

In an important contribution in (Ninness and Hjalmarsson, 2004) it was shown that the asymptotic average variability of a transfer-function estimator at a fixed frequency \( z \in \mathbb{T} \) is given by \( K_{P^*}(z, z) \), i.e., the Cramér-Rao kernel evaluated on its diagonal.

The goal of this section is twofold. In Section 6.1 we demonstrate how the asymptotic average Cramér-Rao kernel can be computed without invoking a parametrization. In Section 6.2 we stress that when doing input design, it is the Cramér-Rao kernel that matters and not the information matrix with respect to a particular parametrization.

6.1 Computation of the Cramér-Rao Kernel

In the preceding sections the information metric quantified the “information” contained in one sample \( x_1 \in X \) where \( X \subset \mathbb{R}^d \) is the sample space. In applications like System Identification an observation consists of a sample vector
\[
x_N = (x_1, \ldots, x_N) \in X^N,
\] (30)
where \( N \) is called the sample-size. The goal of this section will be to quantify the asymptotic average Cramér-Rao kernel, i.e., to quantify the autocovariance of an efficient transfer function estimator.

If one replaces \( x \) in (27) by the sample vector \( x_N \) given by (30) one obtains a sample size dependent information metric. To distinguish between different sample sizes we use the notation \( (\cdot, \cdot)_{N, P^*} \) to denote the information-metric given \( N \) samples.

As an example assume \( \mathcal{P} \) denotes a model set consisting of asymptotically stable transfer-functions, and
\[
x_t = (u_t, y_t) \in X = \mathbb{R}^2 \ \text{s.t.} \ \ y = P^* u + v,
\] (31)
where the noise-process \( v \) and the input-process \( u \) are assumed independent, with spectra given by \( \Phi_v, \Phi_u : T \to \mathbb{R} \) respectively.\(^4\)

A fundamental contribution in (Caines and Ljung, 1979) was to show that, with sample size \( N \to \infty \), the expression \( N^{-1}(\Delta_1, \Delta_2)_{N,P} \) tends, for all \( \Delta_i \in T_P, P \), to a signal-to-noise ratio weighted \( L_2 \)-inner-product

\[
\langle \Delta_1, \Delta_2 \rangle_{P^*} := \frac{1}{2\pi} \int_\pi \Delta_1(e^{j\omega}) \overline{\Delta_2(e^{j\omega})} \Phi_u(e^{j\omega}) \Phi_v(e^{j\omega}) \, d\omega,
\]

(32) which is therefore called the metric which measures the average information per sample. The next theorem shows that there exist cases where it is possible to compute the tangent-space and the Cramér-Rao kernel without using a parametrization. The key advantage of our geometric point of view is that it allows the usage of existing knowledge on tangent spaces of transfer function manifolds (Helmeke and Fuhrmann, 1998) and structured reproducing kernel Hilbert spaces (Grenander and Szego, 1958). In other words, the coordinate-free approach renders it unnecessary to rederive such results in a parametric setup.

**Theorem 6.1.** Let \( \text{Rat}(n) \) denote the \( 2n \)-dimensional manifold of all i) real rational ii) strictly proper, iii) Schur stable transfer functions of McMillan-degree \( n \), i.e.,

\[
\text{Rat}(n) := \left\{ \frac{b}{p} \mid i) b, p \in \mathbb{R}[z] \text{ are coprime s.t.} \right. \]

\[
\left. \quad \text{ii) } \deg(b) < \deg(p) = n, \text{ and} \quad \right. \]

\[
\left. \quad \text{iii) } p(z) = 0 \implies |z| < 1 \right\},
\]

The tangent-space at \( b/p \in \text{Rat}(n) \) is given by

\[
T_{b/p} \text{Rat}(n) := X^p,
\]

(34) where for any non-zero polynomial \( q \) in \( \mathbb{R}[z] \) we define

\[
X^q := \left\{ \frac{f}{q} \mid f \in \mathbb{R}[z] \text{ with } \deg(f) < \deg(q) \right\},
\]

(35) For \( \Phi_u/\Phi_v = 1 \) the Cramér-Rao kernel w.r.t. to the asymptotic average information is given by

\[
\kappa_{b/p}(z, w) = \left( \frac{\varphi(z)\varphi(w)}{\rho(z)\rho(w)} \right)^2 - 1,
\]

(36) for all \( z, w \in T \). For \( \Phi_u/\Phi_v \equiv 1 \) for some constant \( \rho \) the asymptotic Cramér-Rao kernel is \( K_{b/p} = \kappa_{b/p}/\rho \) which, when evaluated on the diagonal (i.e., for \( z = w \)), yields

\[
K_{b/p}(z, z) = \frac{\Phi_u}{\Phi_v} \sum_{i=1}^n \frac{1 - |a_i|^2}{|z - a_i|^2},
\]

(37) where \( p(z) = \prod_{i=1}^n (z - a_i) \).

*Proof.* The fact that (34) holds has been shown in (Helmeke and Fuhrmann, 1998). The fact that (36) is the reproducing kernel of the complexification of \( X^p \), i.e., of the space

\[
X^p := \left\{ \frac{a}{q} \mid a, q \in \mathbb{C}[s], \deg(q) < 2n \right\},
\]

(38) is a well known fact, see e.g. (Ninness and Gustafsson, 1994). By Remark A.1 in Appendix A this implies the reproducing property, i.e., that \( \kappa_{b/p} \) defined in (36) is indeed the asymptotic average Cramér-Rao kernel. The remaining equation (37) follows from (36) by straightforward calculation. □

**Remark 6.1.** Our assumption that the signal to noise ratio \( \rho \) is constant leads to insight as to how the poles of the true system influence the Cramér-Rao kernel. In the general case this dependence is more complicated. Nevertheless one can obtain the Cramér-Rao kernel for the case \( \Phi_u/\Phi_v = \rho \equiv \text{const} \) using Szego’s formula (Grenander and Szego, 1958):\(^2\)

\[
K_{b/p}(z, w) = \frac{1}{\rho(z)\rho(w)} \varphi(z)\varphi(w) - \frac{1}{1 - z^\rho} \varphi(z),
\]

(39) for all \( z, w \in T \), where

\[
\alpha = \frac{1}{2\pi} \int_\pi |\varphi(e^{j\omega})|^2 \frac{\rho(e^{j\omega})}{|\rho(e^{j\omega})|^2} \, d\omega,
\]

(40) and

\[
\varphi(z) := \begin{pmatrix} c_0 & \cdots & c_{-2n+1} & c_{-2n} \\ \vdots & \ddots & \ddots & \vdots \\ c_{2n-1} & \cdots & c_0 & 1 \end{pmatrix},
\]

(41) and \( c_{-2n}, \ldots, c_n \) denote the generalized moments of the signal to noise ratio, i.e.,

\[
c_k = \frac{1}{2\pi} \int_\pi e^{jkw} \frac{\rho(e^{j\omega})}{|\rho(e^{j\omega})|^2} \, d\omega \quad \text{for} \quad k = \pm 1, \ldots, \pm 2n.
\]

6.2 **Autocovariance versus classical covariance matrix**

The importance of a coordinate-free framework for system identification arises from the fact that a priori it is not known which model structure to use. Two common criteria which judge the quality of a given model structure are the bias and the variance. In this section we discuss why covariance matrices alone are not suitable for variance quantification, especially when comparing different model structures. As we shall see, the properties of a model structure are not reflected by covariance matrices but by the corresponding Cramér-Rao kernel.

Assume we excite an unknown stable plant \( P \) at distinct frequencies \( \omega_0 \) and \( \omega_1 \) with sinusoidal inputs of amplitudes \( \sqrt{2\rho_0}, \sqrt{2\rho_1} \) (\( \mu_i > 0 \) respectively) and \( e^{j\omega_2} \notin \mathbb{R} \) for \( k = 1, 2 \). Given the resulting input-output data one may wonder whether one should use the model structure Rat(2), defined in (33), or the model structure FIR(4), i.e., a finite impulse response filter of order four.

For all \( \theta \in \mathbb{R}^4 \), except a small subset of Lebesgue measure zero, there exist unique elements \( G_\theta \) in Rat(2), \( F_\theta \) in FIR(4) such that

\[
\begin{bmatrix} F_\theta(e^{j\omega_0}) \\ F_\theta(e^{j\omega_1}) \\ G_\theta(e^{j\omega_0}) \\ G_\theta(e^{j\omega_1}) \end{bmatrix} = \begin{bmatrix} \theta_1 + j\theta_2 \\ \theta_3 + j\theta_4 \end{bmatrix}.
\]

In other words both model structures can be parametrized by their values \( \theta_1 + j\theta_2 \) and \( \theta_3 + j\theta_4 \) at the interpolation points \( \omega_0 \) and \( \omega_1 \), respectively.

For ease of exposition assume that the noise spectrum is flat, say \( \Phi_v = 1 \). Since the input spectrum \( \Phi_u \) is a point measure at \( \omega_0 \) and \( \omega_1 \), the asymptotic average information matrix \( J_\theta \) at the parameter \( \theta_* \), corresponding to the unknown system.
\[
\begin{bmatrix}
\theta_{1,1} + j \theta_{1,2} \\
\theta_{1,3} + j \theta_{1,4}
\end{bmatrix} =
\begin{bmatrix}
P_\ast(e^{j\omega_0}) \\
P_\ast(e^{j\omega_1})
\end{bmatrix},
\]
can be easily computed from (32) and satisfies
\[
J_{\theta} =
\begin{bmatrix}
\mu_0 \\
\mu_0 \\
\mu_1 \\
\mu_1
\end{bmatrix},
\]
\[
jointly of the true system \(P_\ast\), and independently of whether the model structure is FIR(4) or Rat(2). This is not too surprising if one keeps in mind that, at the input frequencies \(\omega_k \in \{\omega_0, \omega_1\}\), estimators \(\hat{\theta}(e^{j\omega_k})\) and \(\hat{G}(e^{j\omega_k})\), which minimize the prediction error, define asymptotically equivalent estimates \(\hat{P}_\ast(e^{j\omega_k}) \in \mathbb{C}\) (k = 0, 1). Therefore the covariance matrix \(J_{\theta}^{-1}\) offers no information about how the variance differs between the two model structures. One should note that \(\hat{F}\) and \(\hat{G}\) are unbiased estimators of \(F_\ast\) and \(G_\ast\), respectively, and not of \(P_\ast\), in general. For instance, the mean square error of \(\hat{F}\), at a fixed frequency \(\omega\), is given by the expression
\[
\left|\hat{P}_\ast(e^{j\omega}) - \hat{F}_\ast(e^{j\omega})\right|^2 + \mathbb{E}_{P_\ast} \left|\hat{F}_\ast(e^{j\omega}) - \hat{F}(e^{j\omega})\right|^2,
\]
and similarly for \(\hat{G}\). It is clear that bias and variance are model structure dependent. The approximate mean square error of \(\hat{F}\), for large sample size \(N\), is
\[
\left|\hat{P}_\ast(e^{j\omega}) - \hat{F}_\ast(e^{j\omega})\right|^2 + N^{-1} |K_{\text{FIR}(4),\hat{P}_\ast}(e^{j\omega}, e^{j\omega})|,
\]
and similarly for \(\hat{G}\). The Cramér-Rao kernel reflects the dependence of the variance on the model structure, cf. Fig. 1.

In the next section we shall see how the Cramér-Rao kernel function can be used to ensure robust performance specifications.

7. ROBUST PERFORMANCE SPECIFICATIONS

Robust performance specifications naturally arise if one performs application oriented identification (Hjalmarsso, 2009; Gevers, 2005). In the following we shall derive a sufficient condition for an efficient estimator \(\hat{P}\) of \(P_\ast \in \mathcal{P}\) to satisfy the following performance specification:

\[
\alpha \leq \text{Prob}\{x \in X^N : \sup_{z \in B} \left|\hat{P}_\ast(z) - P_\ast(z)\right| \leq \frac{\varepsilon}{N}\}.
\] (42)

In other words the experimental conditions satisfy the performance specification if, with a probability not less than \(\alpha\), we have \(x\) is a good sample, i.e., that the supremum norm of the mismatch \(\hat{P}_\ast - P\) on the frequency band \(B \subseteq \mathbb{T}\) is bounded from above by \(\varepsilon/N\).

**Theorem 7.1.** An asymptotically (in \(N\)) sufficient condition for (42) is given by

\[
\forall z \in B : \quad |K_{P_\ast}(z, z) + |K_{P_\ast}(z, z^{-1})| \leq \frac{2 \cdot \varepsilon}{\chi^2_{d}(d)},
\] (43)

where

- \(d\) is the dimension of the differentiable manifold \(\mathcal{P}\),
- \(\chi^2_{d}(d)\) is the \(\alpha\)-quantile of the Chi-square distribution with \(d\) degrees of freedom,
- \(K_{P_\ast}\) is the asymptotic average Cramér-Rao kernel.

**Proof.** As \(N \rightarrow \infty\) the mismatch \(\hat{P}_\ast - P\) can be approximated with a random, i.e., sample dependent, tangent vector \(\Delta_x \in \mathcal{T}_P\mathcal{P}\) which, by asymptotic normality, can be assumed Gaussian with zero mean and Chi-square distributed squared norm, i.e.,

\[
\left|\Delta_x\right|_P^2 := \langle \Delta_x, \Delta_x \rangle \sim \chi^2(d).
\]

The \(\alpha\)-quantile, denoted by \(r := \chi^2_{\alpha}(d)\), is defined as the inverse of the cumulative distribution of \(\left|\Delta_x\right|_P^2\) evaluated at \(\alpha\). This is equivalent to

\[
\text{Prob}(B_r) = \alpha \quad \text{ with } \quad B_r = \{x \in X : \left|\Delta_x\right|_P^2 \leq r\}.
\]

A result established in (Ivanov et al., 2009) guarantees that

\[
2 \cdot \Delta_x(z) \leq \left|\Delta_x\right|_P^2 \cdot (|K_{P_\ast}(z, z) + |K_{P_\ast}(z, z^{-1})|),
\]
holds for all \(x \in X\) and all \(z \in T\). So condition (43) implies that the event

\[
B_x = \{x \in X : \forall z \in B : \Delta_x(z) \leq \varepsilon\},
\]
is implied by the event \(x \in B_r\). From this it follows that \(B_x\) has a probability greater than or equal to \(\alpha\).

Before we conclude we give an example of the structural insight which can be obtained from Theorem 7.1.

**Example 7.1.** Let \(P \in \mathcal{P}\) with \(\mathcal{P} = \text{Rat}(2)\) denote a plant with a complex conjugate pole at \(a \in \mathbb{C}\) such that

\[
P = (b/p) \quad \text{where} \quad p(z) = (z - a)(z - a^*)\],

where \(|a| < 1\) and \(b, p \in \mathbb{R}[z]\) coprime. Moreover let

\[
\alpha = 0.9, \quad B = \{e^{j\omega} : |\omega| < \beta\} \quad \text{and} \quad \varepsilon = 0.1.
\]

Assume we want to meet the performance specification (42), under the assumption \(\Phi_{\ast} \equiv \text{const}\) and \(\Phi_u \equiv \text{const}\). The dependence of the minimum amount of input power required, \(\Phi_u \geq \Phi_{u,\text{min}}\) depends on the bandwidth \(\beta \in (0, \pi)\) and the pole location \(a \in \mathbb{C}\). This is illustrated in Fig. 2 for constant absolute value and varying phase of \(\alpha\). Let

\[
C_P(z) = \frac{1}{2} |\kappa_P(z, z) + \frac{1}{2} |\kappa_P(z, z^{-1})|,
\]
and \(\kappa_P\) is given by (36). By Theorem 6.1

\[
\Phi_u \cdot \chi^2_{d}(d) \cdot C_P(e^{j\omega}) \leq \varepsilon \cdot \Phi_u \quad \text{for all} \quad \omega \in [0, \beta],
\]
is equivalent to the inequality (43).
In this paper we have established that the autocovariance 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 model structure. Therefore, in the context of prediction error identification of linear time invariant systems, the quantification of the autocovariance of a transfer function estimator can be split 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 a signal to noise ratio weighted $L_2$-inner-product. We have given an example of how to handle such performance specifications for the case where the specification required the supremum norm of the mismatch between estimator and explanatory model to be less than some constant with prescribed probability.

REFERENCES


Appendix A. REPRODUCING KERNELS

Let $H$ denote a linear space over the complex numbers with $\dim H = n < \infty$ endowed with an inner-product $\langle \cdot, \cdot \rangle : H \times H \rightarrow \mathbb{C}$. If the elements of $H$ are functions defined on a common set $\Omega$, then there exists a unique two variable function $K : \Omega \times \Omega \rightarrow \mathbb{C}$ such that $\langle f, K_w \rangle = \langle f(w) \rangle$ with $K_w(z) = K(z, w)$, (A.1) holding for all $z, w \in \Omega$ and all $f \in H$.

One calls $K$ the reproducing kernel of $(H, \langle \cdot, \cdot \rangle)$ and (A.1) the reproducing property. There are various ways to compute $K$ given $H$ and $\langle \cdot, \cdot \rangle$. A general method, known as the Aitken-Berg-Collar Lemma, is given by $K(z, w) = \sum_{i,j=1}^{n} [G^{-1}]_{ij} b_i(z)b_j(w)$, (A.2) where $b_1, \ldots, b_n$ is an arbitrary basis of $H$ and $G_{ij} = \langle b_j, b_i \rangle$ the associated Gramian.

**Remark A.1.** Let $b_1, \ldots, b_n$ denote a basis of $H$ with $\langle b_j, b_i \rangle \in \mathbb{R}$ and let $X = \left\{ \sum_{i=1}^{n} a_i b_i | a \in \mathbb{R}^n \right\}$. (A.3)

Moreover define the bilinear inner-product via $\langle \cdot, \cdot \rangle : X \times X \rightarrow \mathbb{R}$ with $\langle f, g \rangle = \langle f, g \rangle$, (A.4) for all $f, g \in X$. Then (A.1) is equivalent to $K(z, w) = U_w(z) + J_v(z)$, (A.5) with $U_w, V_w \in X$ uniquely determined by $\re f(w) = (f, U_w)$, (A.6) $\im f(w) = -(f, V_w)$, for all $f \in X$. (A.7)