A New Approach for Estimating the Robustness of Parameter Estimates to Measurement Noise

Dhruva V. Raman, James Anderson, Antonis Papachristodoulou

***Abstract*— We consider the nonlinear, grey-box system identification problem. We establish an approximation of the covariance of a parameter estimate in this context, with several attractive theoretical properties. Our approximation is analogous to the inverse Fisher Information matrix, which approximates the covariance through the Cramer-Rao Lower bound. Indeed, it agrees asymptotically with the Cramer-Rao based covariance estimate in the limit of increasing data, where the theoretical assumptions necessary for both methods hold. However, our approximation requires fewer assumptions. In particular, it can be applied when the process is undermodelled, and does not require consideration of either the magnitude or form of the undermodelling. Thus our covariance approximation is relevant when it is known that the physical process cannot be perfectly recreated by any allowable parameterisation of the model structure.**

I. Introduction

Mathematical models are only approximations of the physical processes they attempt to recreate, and, as such, there is a mismatch between dynamics of the model and the true process. This mismatch is known as model error, and is comprised of variance error, caused by noise-corruption of the data, and undermodelling, caused by neglect of process dynamics in the model [7]. White/grey box system identification is concerned with the selection of a nominal model, from a parameterised family of model structures, which minimises estimated model error. A wealth of literature considers quantification of model error induced by the nominally parameterised model, see e.g. [11, 8, 10, 5].

Often the goal of system identification is not just the development of a good nominal model as previously described, but the estimation of certain parameters of physical significance. In this case, a white/grey-box model structure directly incorporating these parameters is derived from physical principles, and the specific numerical values of the nominal parameterisation are important. So too is the uncertainty associated with these values. This uncertainty generally arises due to noise corruption of the observed process data. It can also arise due to structural unidentifiability of the model [1].

DVR is supported by the EPSRC Systems Biology Doctoral Training Centre. JA is supported by a Junior Research Fellowship from St. John's College, Oxford. AP is partially supported through EPSRC projects EP/J010537/1, EP/J012041/1, and EP/M002454/1. All are with the Control Group, Department of Engineering Science, University of Oxford, 17 Parks Road, OX1 3PJ Oxford, United Kingdom.

Here, multiple parameterisations induce identical model dynamics, and are therefore indiscriminable on the basis of observed data.

Generally one assumes existence of an unknown optimal parameterisation, which represents the 'true' values of the estimated parameters, and would be found in the absence of measurement noise and structural unidentifiability. The nominal parameterisation, which is obtained by fitting the model structure to noisy data, is known as an **estimator** of the optimal parameterisation. In particular we consider the maximum likelihood estimator (MLE) in this paper. A common measure of the uncertainty associated with this estimator is then its covariance with respect to the probability distribution of measurement noise. A standard approximation of this covariance is the inverse of the Fisher Information Matrix (FIM) associated with the parameter estimate [7, Ch. 9]. This is meaningful through the Cramer-Rao Lower Bound: under certain regularity conditions the inverse of the FIM is a lower bound on the true covariance of the estimator. A key drawback of this method is that it assumes that there is no undermodelling. In other words, one assumes that the true physical process and the optimally parameterised model have identical dynamics.

In this paper, we derive a new approximation of the covariance of the MLE, valid when considering discrete-time observations of an unknown dynamic process corrupted by Gaussian measurement noise. As with all methods, the approximation becomes more accurate as the signal-to-noise ratio of data decreases. However the source of error in our approximation is different from that in the FIM-based method. Thus the two methods could be considered complementary in that for problems in which one approximation is inaccurate, the other may be accurate. Unlike the FIM, it is mathematically valid even when the process is assumed to be undermodelled. In fact, no quantification of the magnitude or form of undermodelling is required. In this sense it belongs to the field of non-parametric statistics, along with more recognised procedures such as the bootstrap method for estimating confidence intervals (see e.g. [6]). Unlike the bootstrap method, accuracy does not rely on the distribution of datapoints over the performed number of experimental repeats accurately representing the (unknown) true distribution, but only on the sample mean of the datapoints accurately representing the (unknown) true mean. Thus fewer experimental repeats might be necessary for a reasonable approximation. On the other hand, we neglect

higher moments of the MLE distribution, which can be approximated through the bootstrap method. Note that our method is valid for nonlinear models.

II. Problem Formulation

A. Notation

We denote the Lebesgue measure of a set $S \in \mathbb{R}^n$ by $\lambda(S)$. The determinant of a matrix $A \in \mathbb{R}^{n\times n}$ is denoted $|A|$. The Frobenius distance between matrices $A, B \in \mathbb{R}^{n\times n}$ is denoted $\|A - B\|_F$. The relation $\xrightarrow{D}$ denotes convergence in distribution, while $\xrightarrow{P}$ denotes convergence in probability.

B. Problem Statement

Assume that data Y is generated from discrete-time observations of a process corrupted by Gaussian noise. Specifically,

$$Y_i^j \sim \mathcal{N}(g(t_j), \Sigma^j), \quad i \in \{1, \dots, K\}, j \in \{1, \dots, r\}. \quad (1)$$

Here t_j denotes the j^{th} measured time point, and K denotes the number of experimental runs undertaken. Note that data is i.i.d, so the distribution of Y_i^j is invariant with respect to i. Meanwhile, $g : \mathbb{R}^+ \to \mathbb{R}^n$ is an **unknown** deterministic process, and Σ^j is the covariance matrix associated with measurement noise. In the sequel we shall assume, for ease of exposition, that this is a diagonal matrix, although all results of this paper are easily generalised to the case in which this is not true.

We assume that the distribution of data is **modelled** (but not generated) as follows:

$$Y_i^j \hat{\sim} \mathcal{N}(y(t_j, \theta^*), \Sigma^j) \quad i \in \{1, \dots, K\}, j \in \{1, \dots, r\}. \quad (2)$$

Here $y(t_j, \theta^*) \in \mathbb{R}^n$ is the output of a **known** parameterised deterministic model at time $t_j \in \mathbb{R}^+$, with $\theta^* \in \mathbb{R}^q$ being the **unknown**, unique optimal parameterisation. θ^* is defined as the limit of the maximum likelihood estimator (MLE) as $K \to \infty$. The model can also be parameterised by any member θ of the set $\Theta \subseteq \mathbb{R}^q$. Note that we have attached a hat to the $\sim$ symbol in (2). This signifies that the true distribution of the data is provided by (1), and as such (2) may actually be incorrect. Nevertheless $y(t, \theta^*)$ is the best approximation of the unknown process $g(t)$ within the parameterised model structure constructed *a priori*.

In this setup, the system identification problem consists both of finding the parameterisation θ^*, and approximately quantifying the model error, which can be taken as the distance between $y(t, \theta^*)$ and $g(t)$ according to some metric. Given data Y, we take the maximum likelihood estimator $\theta^{est}(Y)$ as our best guess of θ^*, where

$$\theta^{est}(Y) = \arg\min_{\theta\in\Theta} \sum_{i=1}^{K}\sum_{j=1}^{r} \left(Y_i^j - y(t_j, \theta)\right)^T \Sigma^j \left(Y_i^j - y(t_j, \theta)\right). \quad (3)$$

If we take $\hat{\mu}^j = \frac{1}{K}\sum_{i=1}^{K} Y_i^j$, i.e. the vector of sample means of the data at each timepoint, (3) can be simplified to

$$\theta^{est}(\hat{\mu}) = \arg\min_{\theta\in\Theta} \nu(\hat{\mu}, \theta), \text{ where}$$
$$\nu(\hat{\mu}, \theta) = \sum_{j=1}^{r} \left(y(t_j, \theta) - \hat{\mu}^j\right)^T \Sigma^j \left(y(t_j, \theta) - \hat{\mu}^j\right). \quad (4)$$

The sample means are distributed as

$$\hat{\mu}_j \sim \mathcal{N}\left(\mu, \frac{1}{K}\Sigma^j\right), \quad (5)$$

where $\mu \in \mathbb{R}^{n\times r}$ is the concatenation of the uncorrupted process outputs, i.e. $\mu_j = g(t_j) \in \mathbb{R}^n$.

We make some assumptions on the problem:

A1 Consistency:

$$\lim_{K\to\infty} \theta^{est}(\hat{\mu}) = \theta^* \quad \mathcal{P}\ a.s. \quad (6)$$

A2 $\mathcal{C}^2$ Differentiability: $\nabla_\theta y(t, \theta)$ and $\nabla^2_\theta y(t, \theta)$ are defined for all $\theta \in \Theta$, $t \in \mathbb{R}^+$.

A3 Unique nominal parameterisations: Let

$$W = \{x \in \mathbb{R}^{n\times r} : \arg\min_{\theta\in\Theta} \nu(\hat{\mu}, \theta) \text{ is not unique}\}$$

Then $\lambda(W) = 0$.

A4 Let $U = \{x \in \mathbb{R}^{n\times r} : \theta^{est}(x) \notin \mathcal{C}^2\}$. Then $\lambda(U) = 0$.

A3 guarantees that θ^{est} is a well-defined mapping from data space to parameter space. This allows us to formulate A4. Assumption A1 requires technical conditions covered in [7, Ch. 9]. These include structural identifiability of the model at θ^*, which is hard to check in general, but necessary for well-posedness of optimisation protocols solving (3) and a prerequisite to any parameter estimation protocol. The relationship between structural identifiability and consistency is expanded upon at the beginning of Section IV. Assumption A2 is easily fulfilled when $y(t, \theta)$ is the output of an ODE with a differentiable vector field. Note that Assumptions A3 and A4 are not restrictive conditions. To be broken, we would require the existence of a ball of strictly positive radius in $\mathbb{R}^{n\times r}$, on which θ^{est} was nowhere twice differentiable (for A4) or uniquely defined (for A3).

The problem addressed in the paper can now be formulated. We wish to estimate the covariance of $\theta^{est}(\hat{\mu})$, according to the probability distribution induced by the measurement noise. This is denoted $cov(\theta^{est}(\hat{\mu}))$ in the sequel.

III. Results

A. Review of the Cramer-Rao Lower Bound

We now briefly review standard results on the covariance of the MLE as provided in e.g. [7, Ch. 9]. Given a parameterisation θ, the Fisher Information Matrix of θ is defined as

$$I(\theta) = K\sum_{j=1}^{r} \left(\frac{\partial y}{\partial\theta}(t_j, \theta)\right)^T (\Sigma^j)^{-1} \left(\frac{\partial y}{\partial\theta}(t_j, \theta)\right). \quad (7)$$

Let $\mathcal{Y}(T,\theta) \in \mathbb{R}^{n\times r}$ be the matrix obtained by concatenating the vectors: $\{y(t_j,\theta)\}_{j=1}^r$. The matrix $\Sigma \in \mathbb{R}^{nr\times nr}$ is taken as a block diagonal matrix, with the entries $\{\Sigma^j\}_{j=1}^r$ along the diagonal. In this way, we can reformulate $I(\theta)$ as the matrix product:

$$I(\theta) = K[\nabla_\theta \mathcal{Y}(T,\theta)]^T \Sigma^{-1} [\nabla_\theta \mathcal{Y}(T,\theta)]. \tag{8}$$

Suppose temporarily that $g(t_j)$ and $y(t_j,\theta^*)$ are identical for all t_j (i.e. the model structure includes the true process). In this case, under easily fulfilled regularity conditions detailed in [2], the Cramer-Rao lower bound holds. This states that the covariance of any estimator of θ^* is greater than the inverse of the FIM. This includes our MLE $\theta^{est}(\hat{\mu})$, so we have that

$$cov(\theta^{est}(\hat{\mu})) \geq I(\theta^*)^{-1}. \tag{9}$$

The MLE $\theta^{est}(\hat{\mu})$, as opposed to other estimators, also satisfies asymptotic normality: (9) attains equality in the limit of increasing data quantity. Mathematically,

$$\lim_{K\to\infty} cov(\theta^{est}(\hat{\mu})) = I(\theta^*)^{-1} \ \mathcal{P}\ a.s. \tag{10}$$

Note that these results no longer hold when we assume the process is undermodelled, i.e. we drop the assumption that $g(t_j) = y(t_j,\theta^*)$ for all measured timepoints t_j. Furthermore, in practical terms, $I(\theta^{est}(\hat{\mu}))$ is commonly used to approximate $I(\theta^*)$, whose inverse in turn approximates $cov(\theta^{est}(Y))$ (see e.g. [4, 9, 14]), yielding a covariance estimate of variable quality [6].

B. The Limiting Distribution of the Maximum Likelihood Estimator

We now derive new results on the distribution of the MLE in the limit of increasing data quantity. Critically, our results are valid when the process is undermodelled. A method of calculating the covariance of this limiting distribution is then provided.

Lemma 3.1:

$$\theta^{est}(\hat{\mu}) - \theta^{est}(\mu) \xrightarrow[K\to\infty]{D} \mathcal{N}\Big(0, \nabla\theta^{est}(\mu)\frac{\Sigma}{K}\nabla\theta^{est}(\mu)^T\Big). \tag{11}$$

Proof: First note that A1 and A3 together ensure that $\theta^{est}(\mu) = \theta^*$. Taking a Taylor Expansion of $\theta^{est}(\hat{\mu})$ around μ gives

$$\theta^{est}(\hat{\mu}) - \theta^* = [\nabla\theta^{est}(\mu)]^T[\hat{\mu}-\mu] + O([\hat{\mu}-\mu]^T[\hat{\mu}-\mu])$$

The distribution of the first order term then follows from the distribution of the sample mean given in (5). We get

$$\nabla\theta^{est}(\mu)]^T[\hat{\mu}-\mu] \sim \mathcal{N}\Big(0, \nabla\theta^{est}(\mu)\frac{\Sigma}{K}\nabla\theta^{est}(\mu)^T\Big)$$

It remains to show that the high-order terms in the Taylor Expansion are unimportant in the limit of increasing data. The Law of Large Numbers, recalling (5), gives

$$\hat{\mu} - \mu \xrightarrow[K\to\infty]{P} 0.$$

This means that we can apply the multivariate Delta Method (see e.g. [2], p242) to each component $\theta_i^{est}(\hat{\mu})$ of the nominal parameterisation, providing the result. ■

Lemma 3.1 provides a limiting distribution for the nominal parameterisation. We now provide a tractable expression for the covariance term in this distribution.

Theorem 3.1: Let $g : \mathbb{R}^+ \to \mathbb{R}^n$ be any deterministic process producing experimental data according to (1). Let $x \in \mathbb{R}^{n\times r}$ be a random variable such that $x^j \in \mathbb{R}^n$ is distributed according to (1). Suppose that $\theta^{est}(\cdot)$ is $\mathcal{C}^1$ differentiable in an open neighbourhood $N \ni x$, and take $\theta^{est}(x) = \hat{\theta}$. Then, if $|\nabla_\theta^2\nu(x,\hat{\theta})| \neq 0$, we have

$$\nabla\theta^{est}(x) = \Big(\nabla_\theta^2\nu(x,\hat{\theta})\Big)^{-1}\nabla_x\nabla_\theta\nu(x,\hat{\theta}) \tag{12}$$

Moreover, if $\nabla_\theta\nu(x,\hat{\theta})$ at x is $\mathcal{C}^k$ differentiable for $k > 1$, then so too is $\theta^{est}(x)$.

Proof: First order optimality conditions imply that $\nabla_\theta\nu(x,\hat{\theta}) = 0$. The implicit function theorem [12] guarantees the existence of a unique, continuously differentiable function $\psi : \mathbb{R}^{n\times r} \to \mathbb{R}^q$, and an open set $U \ni x$, such that:

$$\nabla_\theta\nu(z,\psi(z)) = 0 \quad \forall z \in U.$$

Without loss of generality we can assume $U \subseteq N$. By first order optimality conditions, we have

$$\nabla_\theta\nu(z,\theta^{est}(z)) = 0 \ \ \forall z \in U.$$

Uniqueness of ψ then implies that $\theta^{est}(z) = \psi(z)$ on U. Their gradients at x must therefore correspond, giving (12). Another consequence of the implicit function theorem is that if $\nabla_\theta\nu(x,\theta)$ is C^k differentiable at x, for some k, then so too is $\psi(x)$, and by extension $\theta^{est}(x)$. ■

Note that nonsingularity of $\nabla_\theta^2\nu(x,\hat{\theta})$ is a key assumption of Theorem 3.1. However, it is not restrictive. To see why, consider the Lebesgue measure of the following set:

$$A = \{x \in \mathbb{R}^{r\times n} : |\nabla_\theta^2\nu(x,\theta^{est}(x))| = 0\}$$

Suppose $\lambda(A) = 0$. This is necessary and sufficient for the probability of drawing data with a singular $\nabla_\theta^2\nu(x,\hat{\theta})$ to be zero. If $\lambda(A)$ were strictly positive, we would require the existence of a $z \in \mathbb{R}^{r\times n}$, $\epsilon > 0$ satisfying

$$\forall y : \|y-z\|_2^2 < \epsilon; \quad D(y) = 0,$$

where $D(y) = |\nabla_\theta^2\nu(y,\theta^{est}(y))|$. Since the mapping $y \to D(y)$ goes from $\mathbb{R}^{n\times r}$ to $\mathbb{R}$, and depends in a complicated way on the characteristics of the model $y(t,\theta)$, one would not expect it to be identically zero on an open set.

C. Estimation of Distribution of Data

We take our estimate of the covariance of θ^{est} as

$$C(\hat{\mu}) = \nabla\theta^{est}(\hat{\mu})\frac{\Sigma}{K}\nabla\theta^{est}(\hat{\mu})^T, \tag{13}$$

as motivated by Lemma 3.1. This can be calculated using Theorem 3.1. We now show that, in the absence of undermodelling (i.e. the special case in which the

Cramer-Rao bound is valid), our covariance estimate agrees asymptotically with the inverse FIM.

Lemma 3.2: Suppose that

$$y(t_j, \theta^*) = g(t_j) \quad \forall j \in \{1, \dots, r\}.$$

Then

$$\lim_{K\to\infty} cov(\theta^{est}(\hat{\mu})) = \lim_{K\to\infty} C(\hat{\mu}) = I(\theta^*)^{-1}, \tag{14}$$

Proof: We have

$$\hat{\mu}^j = y(t_j, \theta^*) + \bar{\mathcal{E}}^j \qquad \bar{\mathcal{E}}^j \sim \mathcal{N}\big(0, \frac{1}{K}\Sigma^j\big).$$

Where A1 holds in an open set surrounding θ^*, we can differentiate (6) to get:

$$\nabla_{\theta^*} \lim_{K\to\infty} \theta^{est}(\hat{\mu}) = \mathbb{I}, \tag{15}$$

since θ^{est} corresponds locally to the identity function. Application of the chain rule to the LHS of (15) consequently gives:

$$\lim_{K\to\infty} \nabla_{\hat{\mu}} \theta^{est}(\hat{\mu}) \nabla_\theta \mathcal{Y}(T, \theta) = \mathbb{I}.$$

Hence, from (8) and (13):

$$\lim_{K\to\infty} C(\hat{\mu}) I(\theta^{est}(\hat{\mu})) = \mathbb{I}$$

■

Unlike $I(\theta^*)^{-1}$, our covariance estimate $C(\hat{\mu})$ in (13) is not a lower bound on the true covariance. However in practical terms this is irrelevant as θ^* is never known: we only gain an estimate $\theta^{est}(\hat{\mu})$, and $I(\theta^{est}(\hat{\mu}))^{-1}$ is not a guaranteed lower bound in the sense of (9).

Error in the respective covariance approximations $C(\hat{\mu})$ and $I(\theta^*)^{-1}$ comes from different sources: Consider the linearisation of our deterministic model,

$$\tilde{y}(t, \theta) = y(t, \theta^*) + \nabla_\theta^T y(t, \theta^*)(\theta - \theta^*).$$

Note that the linearisation does not affect $I(\theta^*)$, calculated according to (7). Furthermore, we now have that $I(\theta) = I(\theta^*)$, $\forall \theta \in \Theta$. This demonstrates how nonlinearities in the dependence of $y(t, \theta)$ on θ couple with the estimation error $\theta^{est}(\hat{\mu}) - \theta^{est}(\mu)$ to induce the approximation error on the FIM: $I(\theta^{est}(\hat{\mu})) - I(\theta^*)$. This error term may magnify when estimating the covariance, as $I(\theta^{est}(\hat{\mu}))$ is furthermore inverted. The consequences of an ill-conditioned FIM on covariance quantification have been explored in [13].

Now consider the original nonlinear model $y(t, \theta)$, and suppose that the parameter estimation routine is itself linearised. Thus:

$$\begin{aligned}\tilde{\theta}^{est}(\hat{\mu}) &= \theta^{est}(\mu) + \nabla_\mu \theta^{est}(\mu)^T(\hat{\mu} - \mu) \\ &= \theta^* + \nabla_\mu \theta^{est}(\mu)^T(\hat{\mu} - \mu) \\ &\Rightarrow \tilde{\theta}^{est}(\hat{\mu}) - \tilde{\theta}^{est}(\mu) \sim \mathcal{N}\Big(0, C(\hat{\mu})\Big)\end{aligned}$$

In other words, the convergence in distribution relation in (11) would be replaced by equality, regardless of K. Furthermore, the covariance estimate would be immune to the sample error $\hat{\mu} - \mu$. So the approximation error in $C(\hat{\mu})$ is induced by the coupling between nonlinearities in the dependence of the parameter estimate on the sample mean, and the sample error $\hat{\mu} - \mu$.

IV. Examples

Assumption A1, as stated previously, is a prerequisite to meaningful parameter estimation. When $y(t_j, \theta^*) = g(t_j)$ $\forall j \in \{1, \dots, r\}$, consistency is equivalent to structural identifiability of the model at θ^*, where structural identifiability is defined as follows:

$$y(t_j, \theta) = y(t_j, \theta^*) \ \forall j \in \{1, \dots, r\} \Rightarrow \theta = \theta^*.$$

This property is hard to verify for nonlinear ODE models [3]. Even then, it ceases to be a sufficient condition for consistency in the case that there is a nonzero residual between the mean data and the trajectory of the optimally parameterised model. This is demonstrated below.

A. Illustrative example

We wish to estimate a process $\dot{g}(t) = Ag(t)$, with

$$A = \begin{bmatrix} -3 & 1 \\ 2 & 1 \end{bmatrix} \qquad g(0) = [4, 5]^T.$$

Noisy process observations are taken according to (1) at arbitrary timepoints $\{t_j\}_{j=1}^r$, with $\Sigma = \mathbb{I}$. Suppose we undermodel this process, by choosing a model of the form: $\dot{x}(t) = \tilde{A}x(t)$, $y(t, \theta) = C(\theta)x(t)$, where

$$\tilde{A} = \begin{bmatrix} A & 0 \\ 0 & G \end{bmatrix} \qquad C(\theta) = \begin{bmatrix} 1 & 0 & \sin(\theta) & -\sin(\theta) \\ 0 & 1 & \cos(\theta) & -\cos(\theta) \end{bmatrix}$$
$$G = \begin{bmatrix} -3.5 & 1.5 \\ 2 & -1 \end{bmatrix} \quad x(0) = [g(0), 1, 1]^T.$$

Note that $\|y(t, \theta) - g(t)\|_2^2$ is invariant to change in θ, for all $t > 0$. In particular, $y(t, \theta')$ can be obtained from $y(t, \theta)$, for $\theta' > \theta$, by a clockwise rotation through the angle $\theta' - \theta$ around $g(t)$ (see Figure 1).

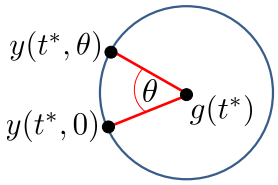

Figure 1. For any $t^* > 0$, the set of parameterisations $y(t^*, \theta)$, for $\theta \in (-\pi, \pi]$ forms a circle with centre $g(t^*)$.

The true probability density function for the sample mean $\hat{\mu}$ of K experiments is Gaussian, with mean μ and variance $\frac{1}{K}\Sigma$ (see (5)). It is therefore invariant over sets of the form: $\Big\{x : (x - \mu)^T \Sigma (x - \mu) = c\Big\}$, for any $c \geq 0$. Since $\Sigma = \mathbb{I}$, we have that $\{y(t, \theta)\}$ is such a set, and the likelihood of any parameter being the MLE is therefore constant over parameter space. Since the distribution of the MLE is uniform over $(-\pi, \pi]$, we can calculate its true covariance as $cov(\theta^{est}(\hat{\mu})) = \frac{1}{3}\pi^2$. Note that this covariance is independent of K, the number of experimental iterations taken.

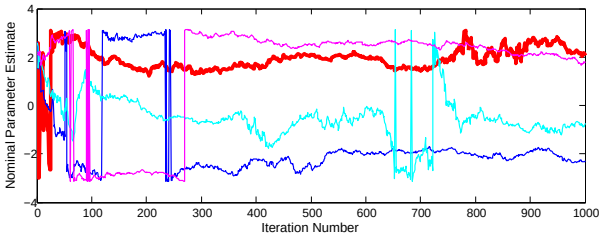

Fig. 2. Plot of $\theta^{est}(\hat{\mu})$. Each iteration alters the sample mean, changing the nominal parameter estimate. This scheme is carried out four times, to emphasise unpredictability. Fig. 3 uses the bold, red, set of iterations.

Assumption A1 (consistency) does not hold, as the covariance does not decay with increasing K. Thus asymptotic covariance quantification of any parameter estimate, either through $C(\hat{\mu})$ or $I(\theta^{est}(\hat{\mu}))^{-1}$, is mathematically invalid. Nevertheless the model is structurally identifiable. Verification/invalidation of Assumption A1 is generally impossible despite being key to much of the theory associated with the field. In this case it is possible only because we know the process dynamics *a priori*. Therefore it is worthwhile to test the fidelity of the covariance approximations $C(\hat{\mu})$ and $I(\theta^{est}(\hat{\mu}))^{-1}$ regardless.

The trajectory of $g(t)$ was simulated multiple times, and data Y was collected by adding measurement noise and storing outputs at the measurement timepoints. At the K^{th} iteration (i.e. after K experimental repeats), an MLE parameter estimate $\hat{\theta}$ was calculated, taking into account the sample mean of all K iterations. A plot of parameter estimate against K is provided in Figure 2. Although the distribution of parameter estimates given K iterations is uniform for any K, as proved previously, this is not true given knowledge of the parameter estimate over $K-1$ iterations. Thus significant autocorrelation is clearly visible. At each iteration K of the experiment, both $I(\theta^{est}(\hat{\mu}))^{-1}$ and $C(\hat{\mu})$ were calculated, and graphs of the estimated covariance for both methods are provided in Fig. 3.

The sample mean $\hat{\mu}$ converges to the true mean μ as iteration number increases. The function $\theta^{est}(\hat{\mu})$ possesses an asymptote at $\hat{\mu}=\mu$, at which point all parameters are optimisers of (3). The gradient $\nabla\theta^{est}(\hat{\mu})$ correspondingly tends to infinity as this asymptote is approached. The estimate $C(\hat{\mu})$, despite being blind to the underlying process dynamics, is predicated on this gradient, and thus flags the estimation problem as ill-conditioned, whereas the FIM-based covariance incorrectly predicts a swiftly decaying covariance (see Fig. 3).

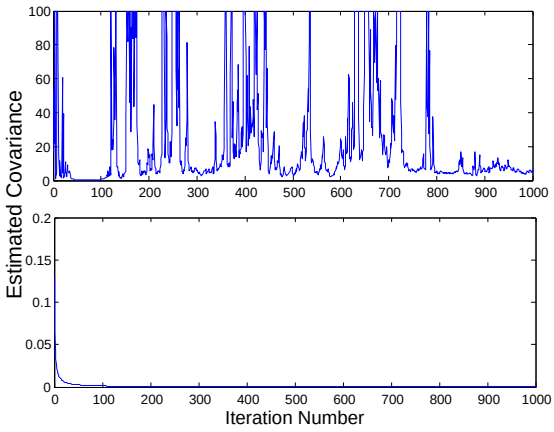

Fig. 3. Estimate of covariance using (13) (TOP), and (7) (BOTTOM), and the sample mean after each iteration. Top figure: any estimates over 100 were taken to be 100 to maintain the scaling of the graph. The maximum estimate was 4.97e5.

B. Chemical reaction network example

We first consider a process that is not undermodelled, i.e. the Cramer-Rao lower bound holds, and compare our covariance estimate $C(\hat{\mu})$ with the FIM-based estimate. They correspond closely. We then alter the process so that it is undermodelled, and again compare estimates. They no longer correspond, and $C(\hat{\mu})$ better approximates the sample covariance of the MLE, which is generated by running multiple estimation routines for different measurement noise realisations. Structural identifiability of the model guarantees A1 when the process is not undermodelled. In the undermodelled case, as discussed, A1 is not generally possible to verify.

Consider the following ODE system:

$$\begin{aligned} &\dot{x}(t,\theta) = A(\theta)x(t) && y(t,\theta) = x(t,\theta) \in \mathbb{R}^3 \\ &A(\theta) = \begin{bmatrix} -0.7-\theta_1 & \theta_2 & 0 \\ \theta_1 & -\theta_2 & 2 \\ 0.7 & 0 & -3 \end{bmatrix} && x(0,\theta) = [50,50,50]^T \\ & && \theta \in \mathbb{R}^2 \end{aligned} \tag{16}$$

which represents, through the law of mass action, the following chemical reaction network structure:

$$x_1 \underset{\theta_2}{\overset{\theta_1}{\rightleftharpoons}} x_2 \qquad x_1 \xrightarrow{0.7} x_3 \qquad x_3 \xrightarrow{2} x_2 \qquad x_3 \xrightarrow{1} \emptyset.$$

We fix $\theta^* = [3,4]$, and take $g(t) = y(t,\theta^*)$. We model the process $g(t)$ through the model structure $y(t,\theta)$, with 'unknown' θ, and attempt to recover θ^* by fitting the model structure to noisy data generated according to (1). $g(t)$ is measured over the vector $T = [0,1,2,\ldots,10]$ of timepoints. We take the measurement noise covariance as $\Sigma = \mathbb{I}$.

Data Y was collected for K process realisations. MLE parameter estimation was performed using the 'greyest' function in MATLAB, to gain an estimate $\theta^{est}(\hat{\mu})$ of θ^*. The covariance estimates $C(\hat{\mu})$ and $I(\theta^{est}(\hat{\mu}))^{-1}$ were calculated, and compared against sample covariance over 1000 parameter estimation with different measurement noise realisations (see Fig. 4, top).

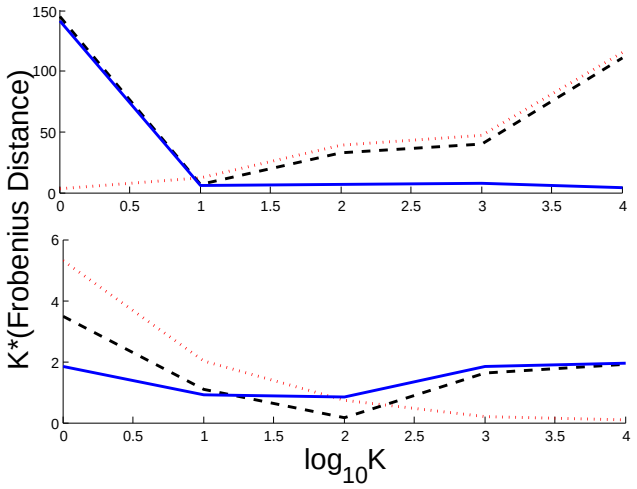

Fig. 4. Sample covariance matrix of θ^{est} taken by running 1000 identifications. Graphs show Frobenius distance between sample covariance and covariance estimators $C(\hat{\mu})$ (red, dotted) and $I(\theta^{est}(\hat{\mu}))^{-1}$ (black, dashed). The blue, solid line represents Frobenius distance between $C(\hat{\mu})$ and $I(\theta^{est}(\hat{\mu}))^{-1}$. Y-axis scaled by K to cancel the effect of absolute decrease in covariance with increasing K. All graphed distances are monotonically decreasing with K in the absence of this scaling. Covariance estimates are taken at increasing integer values of $\log_{10} K$, starting from 0. Model dynamics given by (16) for both figures. TOP: Process dynamics are model dynamics at $\theta^* = [3, 4]$. BOTTOM: Undermodelled case: process dynamics given by (17).

We now alter process dynamics without changing the model structure given in (16), as given below. The process is now undermodelled.

$$\dot{g}(t) = A(\theta^*)g(t) - g_3(t) \qquad \theta^* = [3, 4]^T, \tag{17}$$

$$x_1 \underset{\theta_2^*}{\overset{\theta_1^*}{\rightleftharpoons}} x_2 \qquad x_1 \xrightarrow{0.7} x_3 \qquad x_3 \xrightarrow{2} x_2 \qquad x_3 \xrightarrow{2} \emptyset$$

Here A is as in (16), and $g_3(t)$ represents the third component of the vector $g(t) \in \mathbb{R}^3$. Fig. 4 compares different covariance estimates against sample covariance in this modified case, and demonstrates the superior convergence of $C(\hat{\mu})$.

V. Conclusion

We have introduced a new method of estimating the covariance, with respect to Gaussian measurement noise, of a nominal parameter estimate in the grey-box system identification problem. Our covariance estimator, unlike the traditional inverse Fisher Information Matrix, preserves its mathematical properties when the process is undermodelled. Indeed, no bounds on the form or magnitude of the undermodelling are required. When undermodelling is not a concern, our result agrees asymptotically with the inverse Fisher Information. This was proven and demonstrated by example.

An additional example was provided of a measurement-noise corrupted process, together with a parameterised model, where the model was globally structurally identifiable, yet the parameter estimation problem was not consistent. Unlike the inverse Fisher Information, our covariance estimate flags the ill-conditioned nature of the problem.

References

[1] R. Bellman and K.J. Åström. On Structural Identifiability. *Mathematical Biosciences*, 7:329–339, 1970.

[2] G. Casella and R.L. Berger. *Statistical inference*. Duxbury, 2nd edition, 2002.

[3] O.A. Chis, J.R. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: a critical comparison of methods. *PloS one*, 6(11), 2011.

[4] M.C. Eisenberg and M.A.L. Hayashi. Determining identifiable parameter combinations using subset profiling. *Mathematical biosciences*, 256:116–126, 2014.

[5] G. C. Goodwin and M. E. Salgado. A stochastic embedding approach for quantifying uncertainty in the estimation of restricted complexity models. *International Journal of Adaptive Control and Signal Processing*, 3(4):333–356, 1989.

[6] M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic engineering*, 8(5):447–55, 2006.

[7] L. Ljung. *System identification: Theory for the User*. Birkhauser Boston, 1998.

[8] L. Ljung. Estimating Linear Time-invariant Models of Nonlinear Time-varying Systems. *European Journal of Control*, 7(2-3):203–219, 2001.

[9] H. Miao, X. Xia, A.S. Perelson, and H. Wu. On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM review. Society for Industrial and Applied Mathematics*, 53(1):3–39, 2011.

[10] B. Ninness and G. C. Goodwin. Estimation of model quality. *Automatica*, 31(12):1771–1797, 1995.

[11] A. E. Nordsjö and T. Wigren. On estimation of errors caused by non-linear undermodelling in system identification. *International Journal of Control*, 75(14):1100–1113, 2002.

[12] L. Perko. *Differential equations and dynamical systems*. Springer-Verlag, New York, 1991.

[13] M. Vallisneri. Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. *Physical Review D*, 77(4):042001, 2008.

[14] J. Vanlier, C.A. Tiemann, P.A.J. Hilbers, and N.A.W. van Riel. Parameter uncertainty in biochemical models described by ordinary differential equations. *Mathematical biosciences*, 246(2):305–14, 2013.