12th IFAC International Symposium on
Advanced Control of Chemical Processes
July 14-17, 2024. Toronto, Canada

Sampling Time Design with Misspecified Cramer-Rao Bounds under Input Uncertainty

Ke Wang *,[1] **Hong Yue** *

* *Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XW, UK (e-mails: k.wang@strath.ac.uk, hong.yue@strath.ac.uk).*

Abstract: In the context of parameter estimation, under input uncertainty, the probability distribution function (pdf) of the measurement data mismatches the true pdf of measurement with accurate input. In this scenario, the Cramer-Rao bound (CRB), which is widely used in optimal experimental design, may become an overoptimistic lower bound on parameter estimation error covariance. To tackle this issue of mismatched measurement distribution subject to input uncertainty, in this work, a novel optimal sampling time design is proposed that employs the misspecified Cramer-Rao bound (MCRB), with the aim to collect informative data for high-quality parameter estimation. The MCRB is formed following the Cauchy-Schwarz inequality using the true pdf of the measurement, approximated by the statistics of measurement samples. In the numerical study, large samples from the input uncertainty space are generated and applied to the underlying system model; the outputs are calculated and used to approximate the true measurement pdf. The proposed MCRB-based sampling time design is formulated as a non-convex integer programming optimisation problem solved by a conjugate direction method. Three sampling time designs, the uniform sampling, the CRB-based design and the MCRB-based design, are tested on a benchmark enzyme reaction system model. The results show the necessity and superiority of using MCRB for experimental design under input uncertainty.

Keywords: Input uncertainty, misspecified Cramer-Rao bound (MCRB), optimal experimental design (OED), sampling time design, parameter estimation.

1. INTRODUCTION

A model-building problem includes two key questions: how to obtain the observation data and how to process the data to reach credential conclusions (Lehmann and Casella, 2006)? To answer the former question, the design of experiments, first proposed by Fisher (1937), provides insights on the effective running of experiments and data collection so that the generated and collected data is of the best value for modelling. The experimental design methodology that utilises the prior knowledge of the system's underlying physics is called optimal experimental design (OED) (Barz et al., 2010; Franceschini and Macchietto, 2008) since it often involves an optimisation design based on the Fisher information matrix (FIM).

Collecting measurement data following a well-designed sampling strategy will reduce the experimental efforts and ensure that the collected data contain rich information for modelling (Yu et al., 2018). For the parameter estimation purpose in modelling, rich information means samller parameter estimation error covariance. Lower bounds on parameter estimation error covariance matrix are usually considered as objective functions in OED, especially when the lower bound can be proven tight. The most widely used lower bound for parameter estimation is the Cramer-Rao bound (CRB) (Cramér, 1946; Radhakrishna Rao, 1945).

However, due to imperfect knowledge of the system and imprecise experimental settings in actual implementations, uncertainties are inevitable in both the model and the input signal. This may hamper the reliability of CRB-based design. In our earlier works on sampling time design that incorporate model uncertainty, iterative design strategies (Wang and Yue, 2019) and one-off robust methods are developed using the expected value criterion and the minimax criterion (Wang and Yue, 2020), respectively. For a system with input uncertainty, the sampling time points were selected to minimise the change of output variance caused by input uncertainty, though the design does not consider the information content (Wang and Yue, 2021).

Compared to CRB, the misspecified Cramer-Rao bound (MCRB) is proved to be a tighter lower bound on the error covariance of parameter estimators when the assumed data distribution is different from the true one (Fortunati et al., 2017). The performance of the maximum likelihood estimator (MLE) under the misspecified data model was first studied in Huber et al. (1967) and further developed into a systematic approach in statistics (Vuong, 1986). These fundamental results have been valued in recent applications mainly in signal processing, such as blind channel estimation (Abed-Meraim et al., 2021), MIMO-

[1] Ke Wang is sponsored by China Scholarship Council (CSC)

Radar (Levy-Israel et al., 2022) and direction-of-arrival estimation (Fortunati et al., 2017). Very few works exploit the potential of MCRB in OED. Rosenthal and Tabrikian (2022) proposed a model selection design method in which MCRB is used to assess the performance of linear regression models of different orders. To our knowledge, the MCRB has not been applied to any OEDs for parameter estimation under model or data mismatches.

In this work, a new sampling time design strategy is proposed considering the mismatched measurement pdf from the true one due to input uncertainty, in which the MCRB is used to provide the lower bound for parameter estimation errors. The remaining of the paper is organised as follows: preliminaries are presented in Section 2; in Section 3, the optimal sampling time design is proposed based on the MCRB; the proposed MCRB-based OED is tested with a benchmark enzyme reaction system in Section 4, and compared with two other sampling strategies; conclusions are given in Section 5.

2. PRELIMINARIES

Consider the following state-space model for a dynamic system

$$\dot{\mathbf{X}} = \boldsymbol{f}\left(\mathbf{X};\boldsymbol{\theta}\right),\ \mathbf{X}\left(0\right) = \mathbf{X}_0, \tag{1}$$

where $\boldsymbol{f}\left(\cdot\right)$ are nonlinear functions which are continuous and first-order differentiable; $\mathbf{X} = \left[x_1, \ldots, x_{N_X}\right]^\top \in \mathrm{R}^{N_X}$ is the vector of the state variables; $\mathbf{X}_0 \in \mathrm{R}^{N_X}$ is the initial condition of the states; $\boldsymbol{\theta} \in \mathrm{R}^{N_p}$ is the parameter vector. The output of the system is represented by

$$\mathbf{Y}\left(t\right) = \mathbf{g}\left(\mathbf{X}\left(\boldsymbol{\theta}, t\right)\right) + \boldsymbol{\epsilon}\left(t\right) \tag{2}$$

$\mathbf{g}\left(\cdot\right)$ is considered as a selection function of the states, the measurement noises, $\boldsymbol{\epsilon}\left(t\right) = \left[\epsilon_1\left(t\right), \ldots, \epsilon_{N_Y}\left(t\right)\right]^\top$, are assumed to be independently Gaussian distributed, that is, $\boldsymbol{\epsilon}\left(t\right) \sim N(0, \boldsymbol{\Sigma})$.

2.1 Evaluate Parameter Estimation Quality

A straightforward method to evaluate the quality of an estimator $\hat{\boldsymbol{\theta}}$ is to use the mean square error with respect to (w.r.t.) the true parameter $\boldsymbol{\theta}$:

$$MSE = \mathbb{E}\left\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\right\}. \tag{3}$$

$\mathbb{E}\{\}$ is the expectation function. The best estimator is the one that is unbiased and has the minimum MSE. For an unbiased estimator, the MSE equals to its error covariance matrix $\mathbf{V}(\hat{\boldsymbol{\theta}})$. In fact, for any estimators, $\mathbf{V}(\hat{\boldsymbol{\theta}})$ is the smallest second-order moment, i.e., $\mathbf{V}(\hat{\boldsymbol{\theta}}) \preceq MSE$. Therefore, $\mathbf{V}(\hat{\boldsymbol{\theta}})$ is often used as a measure of parameter estimation quality. Since the analytical form of $\mathbf{V}(\hat{\boldsymbol{\theta}})$ is difficult to obtain for complex dynamic systems, a general lower bound such as CRB serves as an alternative to evaluate the parameter estimation quality.

Assume the experimental data has the assumed pdf, f, the mean of $\hat{\boldsymbol{\theta}}$ is denoted by $\mathbb{E}_f\{\hat{\boldsymbol{\theta}}\} = \boldsymbol{\mu}_f$. Let $\boldsymbol{\zeta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_f$ be the estimation error function. According to the Cauchy-Schwarz inequality, there is

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) \succeq \mathbb{E}_f\left\{\boldsymbol{\zeta}\boldsymbol{\eta}^\top\right\}\mathbb{E}_f^{-1}\left\{\boldsymbol{\eta}\boldsymbol{\eta}^\top\right\}\mathbb{E}_f\left\{\boldsymbol{\eta}\boldsymbol{\zeta}^\top\right\}. \tag{4}$$

where $\boldsymbol{\eta} = \frac{\partial(\ln f)}{\partial\boldsymbol{\theta}}$. Though this inequality holds for any score function $\boldsymbol{\zeta}$, it has been proven that the inequality is most tight if $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ have zero means (Richmond and Horowitz, 2015).

Using this score function and assume $\hat{\boldsymbol{\theta}}$ is unbiased, the inequality (4) is the famous Cramer-Rao inequality:

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) \succeq \mathbf{I}^{-1}(\boldsymbol{\theta}) \triangleq CRB, \tag{5}$$

where

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}_f\left\{\frac{\partial(\ln f)}{\partial\boldsymbol{\theta}}\frac{\partial(\ln f)}{\partial\boldsymbol{\theta}}^\top\right\} = -\,\mathbb{E}_f\left\{\frac{\partial^2 \ln f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right\} \tag{6}$$

is the noted Fisher information matrix (FIM).

2.2 Misspecified Cramer-Rao Bound (MCRB)

Under the input uncertainty, the pdf of measurement data differs from the true measurement pdf with accurate input. Denoting, the true pdf of measurements as p, the expectation of parameter estimator under true data distribution is written as $\boldsymbol{\mu}_p = \mathbb{E}_p\{\hat{\boldsymbol{\theta}}_f(\mathbf{Y})\}$.

Since the covariance inequality (4) holds regardless of the measurement pdf, we can replace the expectation function $\mathbb{E}_f\left\{\cdot\right\}$ by $\mathbb{E}_p\left\{\cdot\right\}$. Reform the error function and the score function in Section 2.1 gives

$$\boldsymbol{\zeta} = \hat{\boldsymbol{\theta}}_f(\mathbf{Y}) - \boldsymbol{\mu}_p, \tag{7}$$

$$\boldsymbol{\eta} = \frac{\partial \ln f(\mathbf{Y} \mid \boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \mathbb{E}_p\left\{\frac{\partial \ln f(\mathbf{Y} \mid \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\}. \tag{8}$$

The MCRB is written as

$$MCRB \triangleq \mathbb{E}_p\left\{\boldsymbol{\zeta}\boldsymbol{\eta}^\top\right\}\mathbb{E}_p^{-1}\left\{\boldsymbol{\eta}\boldsymbol{\eta}^\top\right\}\mathbb{E}_p\left\{\boldsymbol{\eta}\boldsymbol{\zeta}^\top\right\}. \tag{9}$$

When the true measurement data pdf is perfectly fit with the assumed one, i.e., $p = f$, under the condition that $\hat{\boldsymbol{\theta}}_f(\mathbf{Y})$ is unbiased, $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ in (7)-(8) are equivalent to those obained using CRB. Therefore, the MCRB can be seen as a generalisation of CRB. It should be noted that the true measurement distribution model p is difficult to obtain in practice. The Monte-Carlo simulation can be used to get the statistics of output distribution from a large volume of samples (White, 1982).

The MCRB in (9) is explicitly presented by the score function $\boldsymbol{\eta}$ in (8) that is applicable for a specific class of estimators (conditions for these estimators can be found in Richmond and Horowitz (2015)). It is proved in Fortunati et al. (2016) that (9) holds for all misspecified (MS)-unbiased estimators. An estimator is MS-unbiased if and only if

$$\mathbb{E}_p\{\hat{\boldsymbol{\theta}}(\mathbf{Y})\} = \boldsymbol{\theta}_{KL}, \tag{10}$$

where $\boldsymbol{\theta}_{KL}$ is the parameter set that minimises the Kullback-Leibler (KL) divergence between p and f:

$$\boldsymbol{\theta}_{KL} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}_p\left\{\ln(p/f)\right\} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}_p\left\{-\ln f\right\}. \tag{11}$$

3. SAMPLING TIME DESIGN WITH MCRB

3.1 Standard Optimal Sampling Time Design

When the system is nonlinear to parameters, a typical method to deal with the nonlinearity is to apply the first-order Taylor expansion at a specific point, $\boldsymbol{\theta}_0$, and ignore the higher-order terms, which gives

$$\mathbf{X}(\boldsymbol{\theta}) = \mathbf{X}(\boldsymbol{\theta}_0) + \mathbf{S}_0^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{12}$$

where $\mathbf{S}(\boldsymbol{\theta}) = \frac{\partial \mathbf{X}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the parametric sensitivity matrix and $\mathbf{S}_0 = \mathbf{S}(\boldsymbol{\theta}_0)$. Following the additivity of FIM, the inverse of CRB of a nonlinear dynamic system, linearised at $\boldsymbol{\theta}_0$, can be calculated by the following summation:

$$\mathbf{A}(\boldsymbol{\theta}_0) = \sum_t \mathbf{I}(\boldsymbol{\theta}_0, t) = \sum_t \mathbf{S}_0(t)\boldsymbol{\Sigma}^{-1}\mathbf{S}_0^\top(t). \tag{13}$$

To perform the sampling time design, a binary weighting factor, $\boldsymbol{w} = [w_1, \cdots, w_{N_T}]^\top$, is associated to N_T measurable time points as follows

$$\begin{Bmatrix} t_1 & t_2 & \cdots & t_{N_T} \\ w_1 & w_2 & \cdots & w_{N_T} \end{Bmatrix}. \tag{14}$$

Here $w_i = 1$ means the measurement is sampled at t_i and $w_i = 0$ otherwise. Then, the OED for sampling time design can be formulated as

$$\begin{aligned} \boldsymbol{w}^* &= \arg\max_{\boldsymbol{w}} \phi\left(\sum_{i=1}^{N_T} w_i \mathbf{S}_0(t_i)\boldsymbol{\Sigma}^{-1}\mathbf{S}_0^\top(t_i)\right) \\ s.t. &\quad \sum_{i=1}^{N_T} w_i = N \end{aligned} \tag{15}$$

where N is the given total number of sampling time points to be included, and ϕ is a selected scalar function for a matrix. Several scalar functions are commonly used in OED, such as D-optimal (determinant), E-optimal (maximum eigenvalue), and A-optimal (trace). The optimisation problem in (15) is a non-convex integer programming problem that can be solved using conjugate gradient methods.

In this OED, unbiased estimation is assumed, and the CRB is used as the lower bound for parameter estimation error covariance. With input uncertainty, the output/measurement data distribution will be affected, and the generated data series may become non-Gaussian. Moreover, the input uncertainty may cause correlations between the measurement states, losing the assumption of states' independence. Thus, CRB may not be suitable for assessing parameter estimation quality. The MCRB, on the other hand, considers the mismatch between the true pdf and the assumed pdf of measurement data, which could be a better criterion to quantify the lower bound of parameter estimation error.

3.2 Misspecified Maximum Likelihood (MML) Estimator

Before utilising the MCRB in OED, we need to find an MS-unbiased estimator similar to (10) to ensure MCRB is a meaningful estimation quality criterion. The MLE obtained under the assumed pdf model f becomes the MML estimator $\hat{\boldsymbol{\theta}}_{MML}$, when the measurement data pdf is p. It was proven that the MML estimator converges almost surely (a.s.) to $\boldsymbol{\theta}_{KL}$ when the number of data M is sufficiently large (Huber et al., 1967):

$$\hat{\boldsymbol{\theta}}_{MML}(\mathbf{Y}) \xrightarrow[M\to\infty]{a.s.} \boldsymbol{\theta}_{KL} \tag{16}$$

This attribute means that the MML estimator is asymptotic MS-unbiased.

Under suitable regularity conditions, the asymptotic attribute of MCRB for the MML estimator can be obtained (White, 1982), that is,

$$MCRB(\hat{\boldsymbol{\theta}}_{MML}) \xrightarrow[M\to\infty]{a.s.} MCRB(\boldsymbol{\theta}_{KL}) \tag{17}$$

This suggests that the $MCRB(\hat{\boldsymbol{\theta}}_{MML})$ can be used instead of $MCRB(\boldsymbol{\theta}_{KL})$ under sufficient data, and the lower bound of the MML estimator can be calculated even when the true data pdf model p is unknown.

3.3 Optimal Sampling Time Design with MCRB

Denote $p(\mathbf{Y}(t))$ as the true data pdf at time t, the mean is $\boldsymbol{\mu}(t)$ and the variance covariance matrix is $\boldsymbol{\Sigma}_X(t)$. Suppose the data at different time points are distributed independently. The approximated MML estimator and its expectation under the linearised model at $\boldsymbol{\theta}_0$ are:

$$\hat{\boldsymbol{\theta}}_{MML} \approx \boldsymbol{\theta}_0 + \mathbf{A}^{-1}(\boldsymbol{\theta}_0)\sum_t \mathbf{S}_0(t)\boldsymbol{\Sigma}^{-1}(t)[\mathbf{Y}(t) - \mathbf{g}(\mathbf{X}(\boldsymbol{\theta}_0, t))] \tag{18}$$

$$\boldsymbol{\mu}_p = \boldsymbol{\theta}_0 + \mathbf{A}^{-1}(\boldsymbol{\theta}_0)\sum_t \mathbf{S}_0(t)\boldsymbol{\Sigma}^{-1}(t)[\boldsymbol{\mu}(t) - \mathbf{g}(\mathbf{X}(\boldsymbol{\theta}_0, t))] \tag{19}$$

Then the estimation error function $\boldsymbol{\zeta}$ and the score function $\boldsymbol{\eta}$ at $\boldsymbol{\theta}_0$ can be obtained using (7) and (8) to give

$$\boldsymbol{\zeta} = \mathbf{A}^{-1}(\boldsymbol{\theta}_0)\sum_t \mathbf{S}_0(t)\boldsymbol{\Sigma}^{-1}(t)[\mathbf{Y}(t) - \boldsymbol{\mu}(t)], \tag{20}$$

$$\boldsymbol{\eta} = \sum_t \mathbf{S}_0(t)\boldsymbol{\Sigma}^{-1}(t)[\mathbf{Y}(t) - \boldsymbol{\mu}(t)]. \tag{21}$$

Note the assumed independence between measurements at different time points, there is

$$\mathbb{E}_p\left\{\boldsymbol{\zeta}\boldsymbol{\eta}^\top\right\} = \mathbf{A}^{-1}(\boldsymbol{\theta}_0)\mathbf{B}(\boldsymbol{\theta}_0), \tag{22}$$

$$\mathbb{E}_p\left\{\boldsymbol{\eta}\boldsymbol{\eta}^\top\right\} = \mathbf{B}(\boldsymbol{\theta}_0), \tag{23}$$

where $\mathbf{B}(\boldsymbol{\theta}_0) = \sum_t \mathbf{S}_0(t)\boldsymbol{\Sigma}^{-1}(t)\boldsymbol{\Sigma}_X(t)\boldsymbol{\Sigma}^{-1}(t)\mathbf{S}_0^\top(t)$. According to (9), the approximated MCRB for the nonlinear dynamic system can be written as

$$MCRB\left(\hat{\boldsymbol{\theta}}_{MML}\right) \approx \mathbf{A}^{-1}(\boldsymbol{\theta}_0)\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}^{-1}(\boldsymbol{\theta}_0) \tag{24}$$

Note that the MCRB in (24) only serves as a meaningful (tight) bound of MSE between the MML estimator and the true parameter value when $\boldsymbol{\theta}_0$ is reasonably close to the true parameters.

Define

$$\mathbf{A}(\boldsymbol{w}) = \sum_{i=1}^{N_T} w_i \mathbf{I}(\boldsymbol{\theta}_0, t_i), \quad \mathbf{B}(\boldsymbol{w}) = \sum_{i=1}^{N_T} w_i \cdot \mathbf{B}_i \tag{25}$$

where $\mathbf{B}_i = \mathbf{S}_0(t_i)\boldsymbol{\Sigma}^{-1}(t_i)\boldsymbol{\Sigma}_X(t_i)\boldsymbol{\Sigma}^{-1}(t_i)\mathbf{S}_0^\top(t_i)$. The MCRB is written as

$$MCRB(\boldsymbol{w}) = \mathbf{A}^{-1}(\boldsymbol{w})\mathbf{B}(\boldsymbol{w})\mathbf{A}^{-1}(\boldsymbol{w}). \tag{26}$$

Then, the optimal sampling time design with MCRB can be formed as

$$\begin{aligned} \boldsymbol{w}^* &= \arg\min_{\boldsymbol{w}} \ \phi(MCRB(\boldsymbol{w})) \\ s.t. &\quad \mathbf{1}^\top \boldsymbol{w} = N \end{aligned} \tag{27}$$

Since $\mathbf{A}(\boldsymbol{w})$ and $\mathbf{B}(\boldsymbol{w})$ are linear functions of $\boldsymbol{w}$, the optimisation problem (27) can hardly be convex. Therefore, conjugate direction methods, such as Powell's method, are considered to solve this optimisation problem.

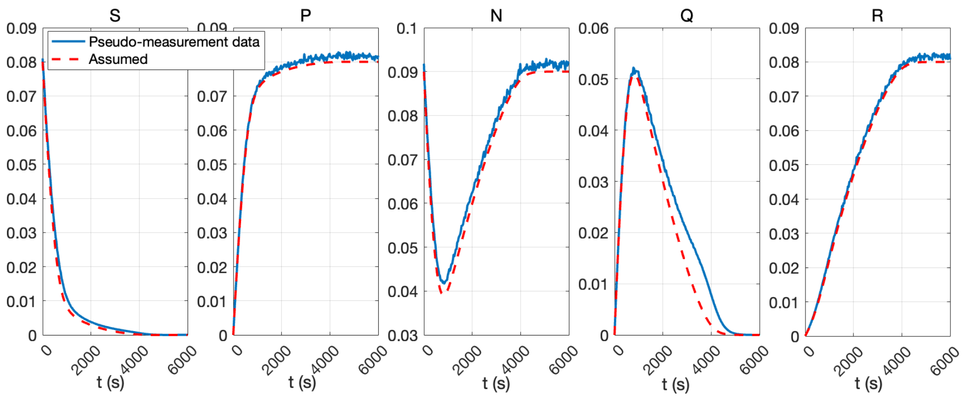

Fig. 1. Standard deviations of the pseudo-measurement data and the assumed data

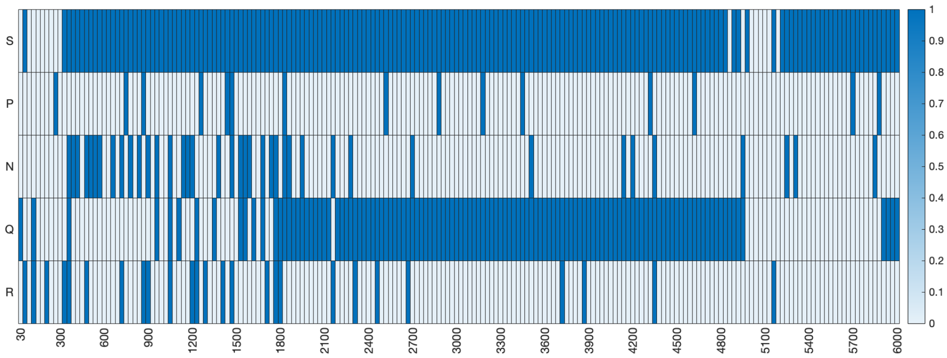

Fig. 2. Chi-squared test with null hypothesis that the pseudo-measurement data follows Gaussian distribution

4. CASE STUDY

The benchmark enzyme reaction system (Yue et al., 2013) represents a typical kinetically controlled synthesis and is a moderate-sized nonlinear dynamic system with ten state variables and eleven parameters. Only three states $[E, S, N]$ have nonzero initial values $[E_0, S_0, N_0]$, which are taken as inputs that can be implemented at the beginning of the reaction process. The non-enzyme states $[S, P, N, Q, R]$ are measurable during the experiments, which are taken as output $\mathbf{Y}$. The kinetic reaction rate vector $\boldsymbol{k} = [k_1, k_{-1}, \cdots, k_6]^\top \in \mathbb{R}^{11}$ includes 11 parameters, in which three of them, k_2, k_{-3}, k_5W, are considered most important according to the parametric sensitivity analysis (Yue et al., 2013). The following assumptions are made for the time sampling design.

- The experimental duration is 6,000 seconds to ensure all state variables reach steady states under step input stimulation.
- The experiment has 200 measurable sampling points, uniformly distributed as [30:30:6000]s, and 20 sampling points need to be selected ($N_T = 200, N = 20$).
- The 5 measurable states are sampled simultaneously.
- The measurement error standard deviation is

$$\sigma_i(t) = \gamma_r x_i(t), \tag{28}$$

where $x_i(t)$ is i-th measurable state at t, γ_r is a relative coefficient, a value of 0.1 is taken for outputs. A small value of 10^{-6} is added to $\sigma_i(t)$ to avoid zero variance.

4.1 Impact of Input Uncertainty on Outputs

The preset inputs are $\mathbf{U}_0 = [E_0, S_0, N_0] = [1.5 \times 10^{-5}, 0.8, 0.9]^\top$. Assume the uncertainties of E_0, S_0, N_0 are independent of each other and follow Gaussian distribution of $N(0, \mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_0$ is a diagonal matrix obtained using (28). The input uncertainty level is set to γ_r at 0.02.

The Monte Carlo method is employed to propagate the input uncertainty to outputs. The key steps are briefed as follows.

(1) Draw 50,000 samples from the uncertainty region of $\mathbf{U}_0 (> \mathbf{0})$ following the Gaussian distribution.
(2) Calculate the model outputs using the prior model and those input samples.
(3) Add measurement noise to the outputs to produce the pseudo-experimental output $\mathbf{Y}$.
(4) Calculate the mean vector and variance matrix of $\mathbf{Y}$.

The results show that the simulated pseudo-measurement data $\mathbf{Y}$ has the same mean as the assumed distribution. The standard deviations of the five states are plotted in Fig. 1, which shows a mismatch from the assumed standard deviations. The Chi-squared test results of the pseudo-experimental data are shown in Fig. 2, in which '0' means accepting the null hypothesis that the data follows the Gaussian distribution, '1' rejecting this hypothesis. It can be seen that the Gaussian assumption of the measurements is also violated. The simulation study results show that the true measurement distribution is mismatched with the assumed one under input uncertainty.

4.2 Sampling Time Design Results

The proposed MCRB-based sampling time design is formed as (27) with the input uncertainty described in the previous section. The D-optimal criterion is adopted as it has been proved to be the most accurate inference on parameter estimations (Ruess et al., 2013). The results of the CRB-based design and the MCRB-based design are shown in Fig. 3 and listed in Table 1, together with the uniform sampling (no design). Compared to the uniform sampling, the selected sampling points of the two OEDs are more gathered in the time regions where sensitivities appear to have larger values. The MCRB-based sampling strategy selects fewer sampling points at the steady states than the CRB-based sampling.

Table 1. Sampling strategies (D-optimal)

Design methods	Sampling strategies (s)
Uniform sampling	[300:300:6000]
CRB-based	[510:30:690, 4140:30:4320, 5100:30:5250]
MCRB-based	**[30, 390:30:600, 3270:30:3600]**

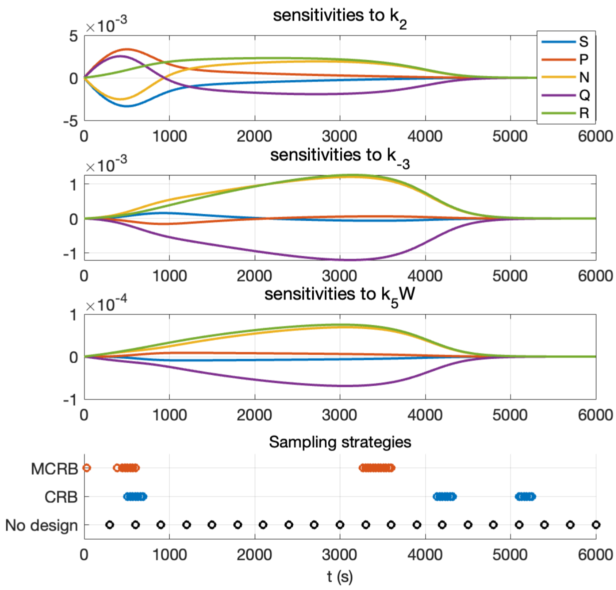

Fig. 3. Relative parameter sensitivities and three sampling strategies (uniform, CRB-based and MCRB-based)

4.3 Performance of Sampling Design Strategies

To compare the performance on parameter estimation quality of the three design strategies in Section 4.2, a number of $4,000$ sets of parameter estimates are obtained by taking the pseudo-measurement data into the MML estimator (18) under the three sampling strategies.

Figure 4 shows the confidence ellipse of these estimations for the three key parameters. It can be seen that compared to the uniform sampling, the parameter estimation errors have been reduced with both CRB-based and MRCB-based designs. Moreover, the proposed MCRB-based design shows smaller confidence regions compared to the standard CRB-based design.

The box-plot is produced to illustrate the locality and spread of the parameter estimates (Fig. 5). The length of the box is the interquartile range (IQR), which is the distance between the first and third quartiles. The median is shown as a horizontal line inside the box. The points are outliers, for they lie beyond 1.5 times of IQR. The whiskers are the locations of the non-outlier maximum and minimum. The numerical results are listed in Table 2, including the range of non-outliers, the median and the mean of the parameter estimates under uniform sampling, CRB-based and MCRB-based sampling designs.

Table 2. Statistics of parameter estimates

Parameter (nominal)	Strategy	range of non-outliers	Median	Mean
k_2 (100)	uniform	28.05	100.8	101.0
	CRB	20.69	100.7	100.8
	MCRB	**19.61**	**100**	**99.90**
k_{-3} (200)	uniform	73.1	197.5	195.6
	CRB	65.6	195.8	194.1
	MCRB	**63.2**	**200.8**	**200.1**
k_5W (5000)	uniform	1673	5004	5003
	CRB	1460	5050	5038
	MCRB	**1160**	**5005**	**4993**

The box-plots in Fig. 5 and the results in Table 2 show that, under the MCRB-based sampling strategy, the parameter estimates have fewer outliers and smaller non-outliers range, and the mean and median are closer to the nominal parameter values compared to the CRB-based design and the uniform sampling. It can be concluded that the MCRB-based OED suits better for a misspecified measurement model. It is found from simulation studies that this conclusion also holds for the E-optimal and A-optimal design criteria.

5. CONCLUSIONS AND FUTURE WORK

Considering the input uncertainty, the assumed pdf of measurements mismatches the pdf of the measurement data under accurate input settings. This mismatch degrades the classical CRB-based OED for parameter estimation. This work proposes a novel sampling time design method that utilises the MCRB instead of the CRB in OED to address the distribution mismatch in measurement data. The developed design method is applied to a moderate-sized benchmark case study system. The simulation results show that this novel design surpasses the classical CRB-based OED and the empirical uniform sampling in that the collected data shows lower bounds for parameter estimation errors. Therefore, MCRB-based OED is a recommended design method for significant input uncertainty.

The MCRB-based OED is developed based on model linearisation around specific parameter values, which mainly works when the nominal model parameters are close to the selected point for linearisation. In addition, other lower bounds except for CRB can be considered for OED with a misspecified model.

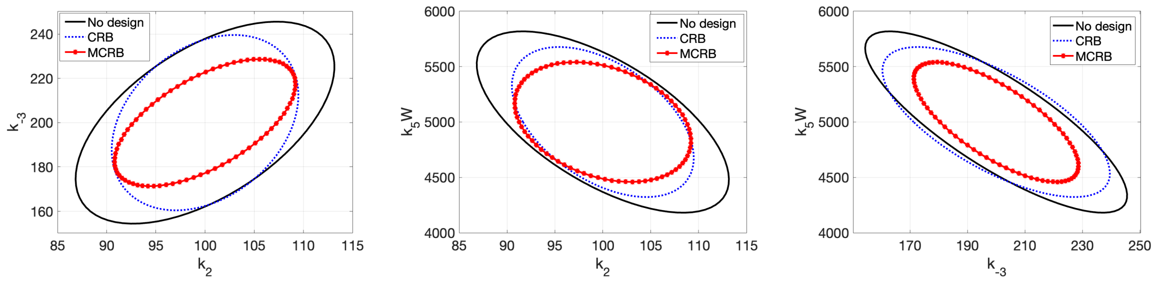

Fig. 4. Confidence ellipse of parameter estimates under three sampling strategies

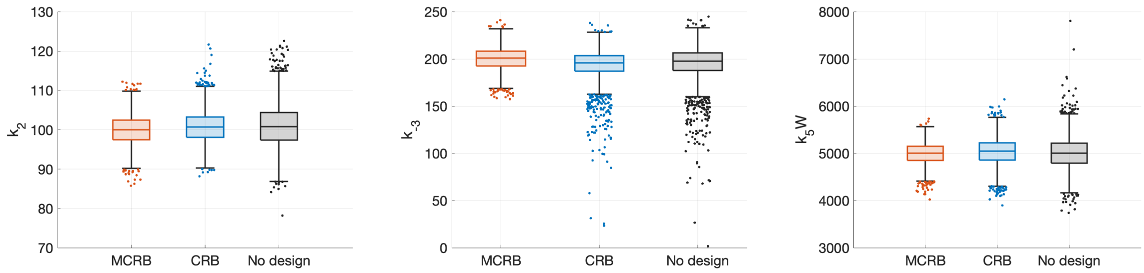

Fig. 5. Box-plot of the parameter estimates under three sampling strategies

REFERENCES

Abed-Meraim, K., Trung, N.L., et al. (2021). Misspecified cramer–rao bounds for blind channel estimation under channel order misspecification. *IEEE Trans. Signal Process.*, 69, 5372–5385.

Barz, T., Arellano-Garcia, H., and Wozny, G. (2010). Handling uncertainty in model-based optimal experimental design. *Ind. Eng. Chem. Res.*, 49(12), 5702–5713.

Cramér, H. (1946). Mathematical methods of statistics. *Princeton U. Press, Princeton*, 500.

Fisher, R.A. (1937). *The Design of Experiments.* 2nd Ed. Oliver & Boyd, Edinburgh & London.

Fortunati, S., Gini, F., and Greco, M.S. (2016). The misspecified cramér-rao bound and its application to scatter matrix estimation in complex elliptically symmetric distributions. *IEEE Trans. Signal Process.*, 64(9), 2387–2399.

Fortunati, S., Gini, F., Greco, M.S., and Richmond, C.D. (2017). Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications. *IEEE Signal Process. Mag.*, 34(6), 142–157.

Franceschini, G. and Macchietto, S. (2008). Model-based design of experiments for parameter precision: State of the art. *Chem. Eng. Sci.*, 63(19), 4846–4872.

Huber, P.J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Math. Stat. & Probab.*, volume 1, 221–233. Berkeley, CA: University of California Press.

Lehmann, E.L. and Casella, G. (2006). *Theory of Point Estimation.* Springer Science & Business Media.

Levy-Israel, M., Bilik, I., and Tabrikian, J. (2022). Misspecified cramér-rao bound for multipath model in mimo radar. In *2022 IEEE 12th Sensor Array and Multichannel Signal Process. Workshop (SAM)*, 81–85. IEEE.

Radhakrishna Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *News Bull. Calcutta Math. Soc.*, 37(3), 81–91.

Richmond, C.D. and Horowitz, L.L. (2015). Parameter bounds on estimation accuracy under model misspecification. *IEEE Trans. Signal Process.*, 63(9), 2263–2278.

Rosenthal, N.E. and Tabrikian, J. (2022). Model selection via misspecified Cramér-rao bound minimization. In *2022 IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP)*, 5762–5766. IEEE.

Ruess, J., Milias-Argeitis, A., and Lygeros, J. (2013). Designing experiments to understand the variability in biochemical reaction networks. *J. R. Soc. Interface*, 10(88), 20130588.

Vuong, Q.H. (1986). Cramér-rao bounds for misspecified models. In *Social Sci. Work. Paper*, volume 652, 1–38. California Institute of Technology.

Wang, K. and Yue, H. (2019). Auto-updating of sampling time redesign for system identification under parameter uncertainty. In *25th Int. Conf. Automat. Comput. (ICAC'19)*, 1–6. IEEE.

Wang, K. and Yue, H. (2020). Robust sampling time designs for parametric uncertain systems. *IFAC-PapersOnLine*, 53(2), 622–627.

Wang, K. and Yue, H. (2021). Robust measurement selection design for experimental systems with input uncertainty. In *26th Int. Conf. Autom. Comput. (ICAC'21)*, 1–6. IEEE.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: J. Econom. Soc.*, 50(1), 1–25.

Yu, H., Yue, H., and Halling, P. (2018). Comprehensive experimental design for chemical engineering processes: A two-layer iterative design approach. *Chem. Eng. Sci.*, 189, 135–153.

Yue, H., Halling, P., and Yu, H. (2013). Model development and optimal experimental design of a kinetically controlled synthesis system. *IFAC Proc. Vol.*, 46(31), 327–332.