2019 American Control Conference (ACC)
Philadelphia, PA, USA, July 10-12, 2019

Transforming Data Across Environments Despite Structural Non-Identifiability

Vipul Singhal[1] and Richard M. Murray[2]

Abstract— **The phenomenon of parameter (structural) non-identifiability can pose significant challenges to the use of parametrized dynamical models. We demonstrate that, for the case of models being used to transform data across environments, it is possible to derive conditions under which the presence of structural non-identifiability does not hinder our modeling objective. We also show that when the non-identifiability has a certain structural feature called (thin) covariation, these conditions are violated, and the transformation methodology must be modified. We demonstrate these results on the problem of correcting batch effects in cell extracts, which are used as rapid prototyping platforms in synthetic biology.**

I. INTRODUCTION

Batch or *environment* specific effects on the measured behavior of a system can be a major confounding factor in data intensive workflows. Often, these environment specific effects may be corrected using calibration measurements from the individual environments. When correcting for the environment specific effects is more complex than a simple scaling of the data, model-based calibration methodologies must be used. These methodologies, however, often depend on estimating model parameters, which in turn are often (structurally) non-identifiable.

In this paper, we derive conditions under which such non-identifiability does not hinder the use of the models for performing the calibrations and the data transformations. Our results rely on the fact that the set of structurally non-identifiable parameters is an equivalence class with respect to the input-output behavior of a system [1], and the data being transformed involves only the input-output behavior of the model, not the trajectories of the full set of state variables.

We begin by defining some notation (Section II) and framing the variability reduction in terms of the so called *data correction problem* (Section III). We then define the *calibration-correction method*, named after a similar method developed to correct wind tunnel variability [2], which solves this problem (Section III). Next, we show that under certain consistency conditions, the presence of parameter non-identifiability does not hinder our methodology (Section IV). We also show that these consistency conditions may be violated when the non-identifiability possesses a certain structural feature called covariation, and end with a modification to the methodology that addresses this phenomenon (Section V).

II. NOTATION AND PRELIMINARY IDEAS

A. Systems, Experiments, Models and Parameters

We consider systems $\mathcal{S} = (\mathcal{E}, \mathcal{P})$ described as a combination of an environment $\mathcal{E}$ and a process $\mathcal{P}$, and define an experiment $\mathcal{H} = (\mathcal{S}, x_0, \overline{y})$ to be the execution of a system under initial conditions x_0 and output measurements $\overline{y}$, where the bar denotes the assumption that experimental data reflects the ground truth. Time dependent inputs may be included without significant change to the results derived in this paper, and are suppressed for simplicity.

The parameter vector θ of a model M associated with a given experiment will be partitioned into *environment specific parameter* (ESP) coordinates $e \in \mathbb{R}^{q_E}$, and *process specific parameter* (PSP) coordinates $p \in \mathbb{R}^{q_P}$. We do not restrict these parameters to be in the positive orthant, since any positive parameters may be log transformed.

The partition of $\theta = (e, p)$ into ESPs and PSPs may be made using the following guidelines: ESPs are parameters associated primarily with components that are present in the system regardless of the the process implemented. In the biological example discussed in this paper, these components are biochemical species, like enzymes and DNA. PSPs are parameters associated with components that may no longer be present in the system when the process is changed.

Experiments are modeled using initialized parametrized models with the equations of the general form

$$\begin{aligned} \dot{x} &= f(x, \theta), \\ y(\theta, x_0) &= h(x, \theta), \qquad x(0) = x_0(\theta). \end{aligned} \tag{1}$$

Here $x, x_0 \in \mathbb{R}^n_+$, the solutions are assumed to exist for all $t \geq 0$, the parameter vector symbol is $\theta = (e, p) \in \Omega \subseteq \mathbb{R}^{q_E + q_P}$, where Ω is the set of all parameter values of interest. The output is denoted $y(\theta, x_0) \in \mathbb{R}^r$. The functions f and h are assumed to be analytic vector fields with respect to x in some neighborhood of any attainable x, and time dependence of the vector fields can be modeled by including t in the state variables [3]. We will use the shorthand $y(\theta, x_0) = M(\theta, x_0)$ to refer to the model in Equation (1), and will often suppress arguments

This work was supported by SBIR-STTR grant W911NF-16-P-0003 and AFOSR grant FA9550-14-1-0060

[1]Vipul Singhal is with the Computation and Neural Systems option, California Institute of Technology, USA; and the Genome Institute of Singapore, Singapore. vsinghal@caltech.edu, vipul_singhal@gis.a-star.edu.sg

[2]Richard M. Murray is with the Department of Control and Dynamical Systems, California Institute of Technology, Pasadena CA 91125, USA. murray@cds.caltech.edu

such as x_0 for brevity. We will sometimes replace (θ) with (e,p), as in stating $y(e,p) = M(e,p)$ or just $M(e,p)$. We will use the hat symbol ($\hat{}$) to denote an estimated parameter value ($\hat{\theta}$, for instance), or a simulated model trajectory, $\hat{y}$. The tilde ($\tilde{}$) over parameter symbols is reserved for miscellaneous purposes, particularly in proofs.

Example 1. Consider a simple model of protein production, given by the chemical equation

$$\text{dna} + \text{enz} \underset{k_r}{\overset{k_f}{\rightleftharpoons}} \text{dna:enz} \xrightarrow{k_c} \text{dna} + \text{enz} + \text{prot}. \tag{2}$$

The corresponding ODE model is given by

$$\begin{aligned} \frac{\mathrm{d}[\text{enz}]}{\mathrm{d}t} &= -k_f[\text{enz}][\text{dna}] + (k_r + k_c)[\text{dna:enz}] \\ \frac{\mathrm{d}[\text{prot}]}{\mathrm{d}t} &= k_c[\text{dna:enz}] \\ \frac{\mathrm{d}[\text{dna:enz}]}{\mathrm{d}t} &= -\frac{\mathrm{d}[\text{dna}]}{\mathrm{d}t} = -\frac{\mathrm{d}[\text{enz}]}{\mathrm{d}t} \end{aligned} \tag{3}$$

and initial conditions $[\text{enz}](0) = \text{enz}_0$, $[\text{dna}](0) = \text{dna}_0$, $[\text{dna:enz}](0) = 0$ and $[\text{prot}](0) = 0$.

Here, we may choose to define the initial enzyme concentration and the reaction catalysis rate k_c as the environment specific parameters, and the remaining parameters (k_f and k_r) as the process specific parameters. This partition will, in general, be application domain specific, and require expert knowledge and empirical trial-and-error. △

B. The Model Universe and Model Correctness Assumptions

Our analytical results will be stated and proved in a virtual *model universe*, where artificial data $\overline{y}$ are generated using nominal models $\overline{M}$ with known nominal parameter values $\overline{\theta}$. I.e., in the model universe, we identify $\mathcal{H} = (\mathcal{S}, x_0, \overline{y})$ with $\overline{y} = \overline{M}(\overline{\theta}, x_0)$.

We will also make a *model correctness* assumption, denoted $M = \overline{M}$, which states that the models we use to estimate parameters from the data are the very models used to generate the data. The model correctness assumption allows us to isolate the interaction of non-identifiability with our method. Issues associated with model correctness or the use of approximate models (that often arise due to model order reduction) have implications on the practical application of these results, and form an important direction for future work. Furthermore, we emphasize that the nominal parameter points generate nominal trajectories, which can in turn be used to compute the sets of output indistinguishable parameters, of which the nominal parameters are members.

In this paper, we will use the model universe assumption to refer to both assumptions.

C. Parameter Non-Identifiability

In this section, we follow Walter and Lecourtier [3] in defining the notion of structural non-identifiability.

Definition 1 (Output-Indistinguishability)**.** Let $M(\theta_A)$ be a parametrized model, and let $M(\theta_B)$ be a model with the same structure. $M(\theta_A)$ and $M(\theta_B)$ are said to be *output-indistinguishable* if

$$\begin{aligned} &\theta_A,\ \theta_B \in \Omega, \\ &y(\theta_A, x_0) = y(\theta_B, x_0) \quad \forall t \geq 0,\ \forall x_0 \in \mathbb{R}^n_+. \end{aligned} \tag{4}$$

Definition 2 (Structural Global Identifiability (parameter))**.** The i^{th} coordinate of θ_A, denoted $\theta_{A,i}$, is *structurally globally identifiable* (SGI) if for almost any $\theta_A \in \Omega$, Equation (4) has a unique solution for $\theta_{B,i}$.

Intuitively, for an SGI coordinate, output indistinguishable trajectories almost always lead to a unique estimate of the coordinate.

Definition 3 (Structural Global Identifiability (model))**.** The model $M(\theta)$ is called *structurally globally identifiable* (SGI) if all its parameters θ_i, for $i = 1, 2, \ldots, q_E + q_P$, are SGI.

In the absence of global structural identifiability, multiple points in the parameter space give rise to the same output behavior. In biological applications for example, this situation tends to be common due to a limited number of measurements and a large number of state variables [4]. Our main goal is to demonstrate that it is not always necessary to achieve global identifiability for every parameter to achieve a modeling objective such as data transformation. To this end, we shall consider models with non-SGI parameters, and thus allow e and p to exist in *sets of output-indistinguishable parameters*, denoted by E and P respectively.

III. THE CALIBRATION-CORRECTION METHODOLOGY

In this section, we frame the problem of transforming data across environments as the so called data correction problem. We also define a stepwise methodology, which we call the calibration-correction method, that attempts to solve this problem.

Consider two environments, a *reference environment* ($\mathcal{E}_1$), and a *candidate environment* ($\mathcal{E}_2$). Let $\mathcal{H}_{i,\text{cal}}$ (resp. $\mathcal{H}_{i,\text{test}}$) be an experiment performed with a *calibration process* $\mathcal{P}_{\text{cal}}$ (resp. *test process* $\mathcal{P}_{\text{test}}$) in the environment $\mathcal{E}_i$. Assume that we may pick models $M_{\text{cal}}(\theta_{\text{cal},i})$ and $M_{\text{test}}(\theta_{\text{test},i})$ corresponding to $\mathcal{H}_{i,\text{cal}}$ and $\mathcal{H}_{i,\text{test}}$ (respectively), as long as the models are at the same level of detail (see Remark 1 and Example 2). Then, solving the *data correction problem* (DCP) involves taking the tuple $(M_{\text{cal}}, M_{\text{test}}, \overline{y}_{1,\text{cal}}, \overline{y}_{2,\text{cal}}, \overline{y}_{2,\text{test}})$ and returning a trajectory $\hat{y}_{1,\text{test}}$, such that $\hat{y}_{1,\text{test}} = \overline{y}_{1,\text{test}}$.

Remark 1. Two models are at the same level of detail if, whenever some mechanism is a part of both models, it has the same mathematical expressions describing it in each model. The reason this requirement is present is that we will be using values of parameters estimated using one model in the other when we attempt to solve the DCP. This also raises the interesting possibility of using models at different levels of detail, as long as one model can be (model order) reduced to the other. Then, it might

be possible to estimate the parameters in one model, and transform them appropriately before using them in the other model. This is left as a future direction of investigation. ◇

Example 2. Consider a data correction problem encountered in the field of synthetic biology. Here, cell extracts are used as prototyping platforms for studying the dynamics of genetic circuits (networks of interacting DNA, RNA, proteins, and other elements from molecular biology), just like wind tunnels are used to prototype airplane designs. These cell extracts must be produced in batches by extracting the gene expression machinery from lysed bacteria, a process that has considerable batch dependent variability. The transformation of a test circuit (process) data from a candidate extract (environment) to a reference extract can be stated as a data correction problem.

For our example, we set the calibration experiment to be the simple protein expression system shown in Example 1, and the test experiment to comprise a transcription factor mediated repression of a reporter protein (such as green fluorescent protein, GFP), modeled as

$$\begin{aligned}
\mathrm{dna}_{\mathrm{rep}} + \mathrm{enz} &\underset{k_{r,rep}}{\overset{k_{f,rep}}{\rightleftharpoons}} \mathrm{dna}_{\mathrm{rep}}{:}\mathrm{enz} \xrightarrow{k_c} \mathrm{dna}_{\mathrm{rep}} + \mathrm{enz} + \mathrm{rep},\\
\mathrm{dna}_{\mathrm{prot}} + \mathrm{enz} &\underset{k_{r,prot}}{\overset{k_{f,prot}}{\rightleftharpoons}} \mathrm{dna}_{\mathrm{prot}}{:}\mathrm{enz} \xrightarrow{k_c} \mathrm{dna}_{\mathrm{prot}} + \mathrm{enz} + \mathrm{prot},\\
2\mathrm{rep} &\underset{k_{r,dim}}{\overset{k_{f,dim}}{\rightleftharpoons}} \mathrm{rep}_2,\\
\mathrm{dna}_{\mathrm{prot}} + \mathrm{rep}_2 &\underset{k_{r,seq}}{\overset{k_{f,seq}}{\rightleftharpoons}} \mathrm{dna}_{\mathrm{prot}}{:}\mathrm{rep}_2,
\end{aligned} \tag{5}$$

where rep and prot are the repressor and reporter proteins respectively, and are produced by the same mechanism as in Example 1. The DNA species, the bound complex species, and the binding rates are now labeled by appropriate subscripts. The third reaction models the dimerization of the repressor, and the fourth reaction models the sequestration (i.e., repression) of the reporter DNA. Here, the ESPs are the same as the ones in model of the calibration experiment in Example 1: the protein production catalysis rate k_c, and the initial concentration of the enzyme, enz_0. The remaining parameters are the PSPs: $k_{f,rep}$, $k_{r,rep}$, $k_{f,prot}$, $k_{r,prot}$, $k_{f,dim}$, $k_{r,dim}$, $k_{f,seq}$ and $k_{r,seq}$. The corresponding ODEs are created in a similar fashion to those in Example 1.

This example also illustrates what we mean by models at the *same level of detail*. The production of protein from DNA is modeled as a single step enzymatic reaction in both the calibration process of Example 1 and the test process in this example. This allows for the estimates of the ESP parameters (k_c and enz_0) from the calibration experiment to be used with the test experiment data, as will be required when we attempt to solve the data correction problem.

With these definitions, the data correction problem is illustrated in Figure 1. Here, experimental data corresponding to the calibration and test gene circuits in two different *E. coli* extracts is shown (at various initial concentrations of the protein coding DNA). The desired data transformation involves transforming the test circuit data from the candidate extract ($\mathcal{E}_2$) to the reference extract ($\mathcal{E}_1$), and is shown by the arrow. △

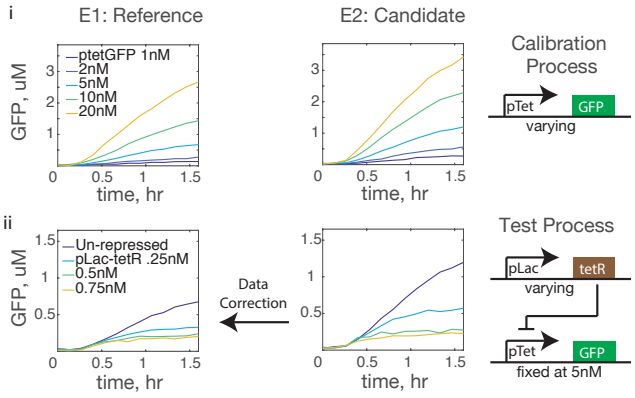

Fig. 1. The data correction problem involves the transformation of the behavior of a *test* process (ii, a tetR repression circuit here), from a *candidate* environment to a *reference* environment. We have the freedom to design and implement a set of *calibration* experiments ($\mathcal{H}_{i,\mathrm{cal}}$, $i = 1, 2$) on the two environments (i, constitutive expression of GFP here), and collect the resulting data ($\overline{y}_{1,\mathrm{cal}}$ and $\overline{y}_{2,\mathrm{cal}}$). The data was collected in TX-TL, using a variation of the protocols found in [5].

Remark 2. In general, the DCP will only be solvable in the model universe, where $\overline{y}_{i,\mathrm{cal}} \triangleq \overline{M}_{\mathrm{cal}}(\overline{e}_i, \overline{p}_{\mathrm{cal}})$ and $\overline{y}_{i,\mathrm{test}} \triangleq \overline{M}_{\mathrm{test}}(\overline{e}_i, \overline{p}_{\mathrm{test}})$, $i = 1, 2$. ◇

Remark 3. With real data, the equality $\hat{y}_{1,\mathrm{test}} = \overline{y}_{1,\mathrm{test}}$ must be replaced with the approximate equality $\hat{y}_{1,\mathrm{test}} \approx \overline{y}_{1,\mathrm{test}}$, defined in some sense. For instance, we may require that $d(\overline{y}_{1,\mathrm{test}}, \hat{y}_{1,\mathrm{test}}) < \frac{1}{2} d(\overline{y}_{1,\mathrm{test}}, \overline{y}_{2,\mathrm{test}})$, where d is an appropriate metric. ◇

Next, define the set valued parameter identification operator that will be used for studying the effect of non-identifiability on our methodology.

Definition 4 (Parameter Identification)**.** Let the set Γ_θ be the set of all pairs $(y, M(\theta))$ for which there exists a parameter $\hat{\theta} \in \Omega$ such that $y = M(\hat{\theta})$. Let $\mathcal{P}(\Omega)$ be the power set of Ω. We define the *parameter identification* of the θ coordinates of the model M as an operation $\mathrm{ID}_\theta : \Gamma_\theta \to \mathcal{P}(\Omega)$, with $\mathrm{ID}_\theta(y, M(\theta)) = \{\hat{\theta} \in \Omega \mid y = M(\hat{\theta})\}$

In the definition above, we have explicitly included θ as a subscript to ID and Γ. This is useful because we also allow for methods of identifying parameters only over some subset of parameter coordinates. The first such method (over p, suppose) is to identify values over all the parameter coordinates, and then to project the resulting set down to the coordinates of interest. This will be denoted by $\mathrm{proj}_p\, \mathrm{ID}_\theta(y, M)$, where $\theta = (e, p)$, and proj is the projection operator. The second method involves a conditional version of the parameter estimation operation, and is defined as follows.

Definition 5 (Conditional Parameter Identification)**.** Consider the partition $\theta = (\theta_a, \theta_b) \in \Omega \subset \mathbb{R}^{q_a} \times \mathbb{R}^{q_b}$. Let $\Gamma_{\theta_a|\theta_b=\tilde{\theta}_b} \triangleq (y, M) \mid \exists \theta_a : y = M(\theta_a, \tilde{\theta}_b)$. Then, we define the *conditional ID* operator as $\mathrm{ID}_{\theta_a|\theta_b=\tilde{\theta}_b} : \Gamma_{\theta_a|\theta_b=\tilde{\theta}_b} \to \mathcal{P}(\mathrm{proj}_{\theta_a} \Omega)$, with $\mathrm{ID}_{\theta_a|\theta_b=\tilde{\theta}_b}(y, M(\theta_a, \theta_b)) = \{\hat{\theta}_a \in \mathrm{proj}_{\theta_a} \Omega \mid y = M(\hat{\theta}_a, \tilde{\theta}_b)\}$.

Remark 4. We unclutter the notation by abbreviating $\mathrm{ID}_{\theta_a|\theta_b=\tilde{\theta}_b}(y, M(\theta_a, \theta_a))$, to $\mathrm{ID}_{\theta_a}(y, M(\theta_a, \tilde{\theta}_b))$, and $\Gamma_{\theta_a|\theta_b=\tilde{\theta}_b}$ to Γ_{θ_a}. ◇

Next, we define the calibration-correction method as a sequence of steps involving parameter identification and prediction. Along with stating each step of the method in terms of single parameter points or trajectories, we also give descriptions of the sets of all such points and trajectories. The definitions of these sets allow for the investigation of how structural non-identifiability of parameters affects this method's ability to solve the DCP.

Briefly, the method involves the following. First, calibration experiment data is fit to corresponding models to find the ESPs for each environment. Next, the ESPs for the candidate environments are fixed in the test model, and the PSPs for this model are estimated using corresponding data. Finally, a corrected trajectory is generated by simulating the test model using these PSPs, along with the ESPs for the reference environment.

Definition 6 (The Calibration-Correction Method)**.** Consider the DCP in the model universe. We define the *calibration-correction method* as a sequence of steps that takes the tuple $(M_{\mathrm{cal}}, M_{\mathrm{test}}, \overline{y}_{1,\mathrm{cal}}, \overline{y}_{2,\mathrm{cal}}, \overline{y}_{2,\mathrm{test}})$ and returns a prediction $\hat{y}_{1,\mathrm{test}}$. The steps are:

1) *Calibration Step.* Find $\hat{e}_{1,\mathrm{cal}}$ and $\hat{e}_{2,\mathrm{cal}}$ such that $(\hat{e}_{1,\mathrm{cal}}, \hat{e}_{2,\mathrm{cal}}, \hat{p}_{\mathrm{cal}})$ satisfies $\overline{y}_{1,\mathrm{cal}} = M_{\mathrm{cal}}(\hat{e}_{1,\mathrm{cal}}, \hat{p}_{\mathrm{cal}})$ and $\overline{y}_{2,\mathrm{cal}} = M_{\mathrm{cal}}(\hat{e}_{2,\mathrm{cal}}, \hat{p}_{\mathrm{cal}})$ for some $\hat{p}_{\mathrm{cal}}$. The sets of all such ESP points are found by projecting the set

$$\tilde{\Theta}_{\mathrm{cal}} \triangleq \left\{ (e_1, e_2, p) \,\middle|\, \overline{y}_{i,\mathrm{cal}} = M_{\mathrm{cal}}(e_i, p),\ i = 1, 2 \right\},$$

onto the corresponding ESP coordinates:

$$E_{i,\mathrm{cal}} \triangleq \mathrm{proj}_{e_i} \tilde{\Theta}_{\mathrm{cal}}, \qquad i = 1, 2. \tag{6}$$

2) *Correction Step One.* Identify $\hat{p}_{2,\mathrm{test}}$ such that $\overline{y}_{2,\mathrm{test}} = M_{\mathrm{test}}(\hat{e}_{2,\mathrm{cal}}, \hat{p}_{2,\mathrm{test}})$. The set of all such points is

$$P'_{2,\mathrm{test}} \triangleq \bigcup_{\hat{e} \in E_{2,\mathrm{cal}}} \mathrm{ID}_{p|e=\hat{e}} \left(\overline{y}_{2,\mathrm{test}}, M_{\mathrm{test}}(e, p) \right). \tag{7}$$

3) *Correction Step Two.* Generate the prediction $\hat{y}_{1,\mathrm{test}} \triangleq M_{\mathrm{test}}(\hat{e}_{1,\mathrm{cal}}, \hat{p}_{2,\mathrm{test}})$. Note that the set of all such predictions is

$$Y_1 \triangleq \bigcup_{\hat{e} \in E_{1,\mathrm{cal}}} \bigcup_{\hat{p} \in P'_{2,\mathrm{test}}} \hat{y}_1(\hat{e}, \hat{p}), \tag{8}$$

with individual predictions $\hat{y}_1(\hat{e}, \hat{p}) \triangleq M_{\mathrm{test}}(\hat{e}, \hat{p})$.

Remark 5. The version of the calibration step defined above is straightforward to implement computationally. It involves a single estimation step, followed by projections. It is also possible to give an equivalent definition that involves estimating the parameters for the two environment separately, which can be useful for reducing the dimension of the spaces over which the parameter estimation must be performed [6]. ◇

Remark 6. Note that the set $P'_{2,\mathrm{test}}$ is a subset of the larger set $P_{2,\mathrm{test}} \triangleq \mathrm{proj}_p \mathrm{ID}_\theta(\overline{y}_{2,\mathrm{test}}, M_{\mathrm{test}})$. Indeed, $P'_{2,\mathrm{test}}$ is obtained from $P_{2,\mathrm{test}}$ by only keeping the points whose corresponding e coordinate values are in the calibration set $E_{2,\mathrm{cal}}$. We use $P'_{2,\mathrm{test}}$ because in the first correction step, we identify p only after fixing the value of e to an arbitrary point within $E_{2,\mathrm{cal}}$. ◇

Example 3. We can perform the calibration-correction method on the data correction problem described in Example 2. The calibration step involves fitting the calibration circuit model to its data, and obtaining the ESP values (k_c and enz_0). The model fits and the MCMC posterior distributions are shown in Figure 2i and 2ii (see the Discussion section (Section VI) for some justification of MCMC as a parameter set identification method). The first three columns of Figure 2iii show the test circuit data to be corrected, the estimation of the process (circuit) specific parameters (correction step 1), and the corrected trajectories (correction step 2) respectively. The final column shows the results of an improvement to the method that we call PSP fixing, and will be discussed when we discuss parameter covariation. △

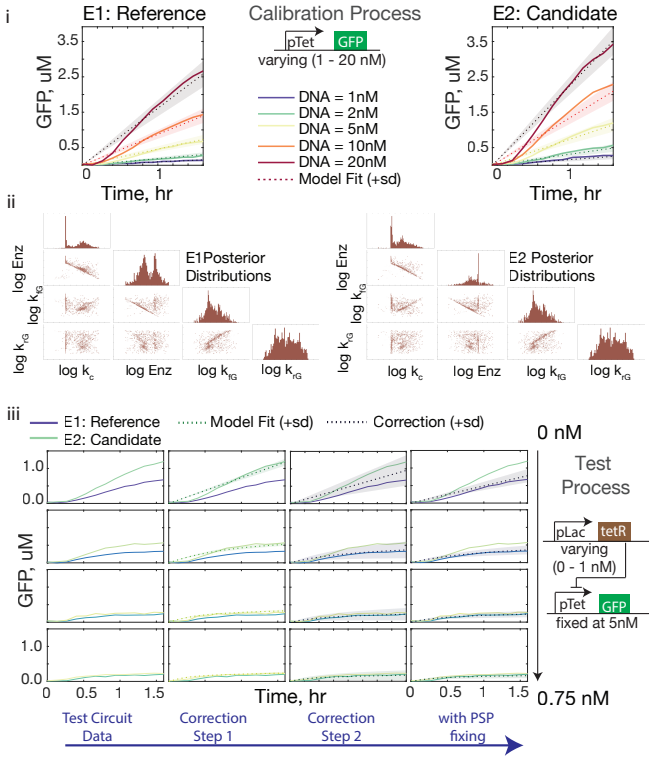

Fig. 2. The calibration-correction method applied to the data correction problem, as described in Examples 2 and 3. MCMC done using [7].

Remark 7. We can define two *failure conditions* for the calibration-correction method that will be useful in deriving the main theoretical results of this chapter. Both the conditions must be avoided for the calibration-correction method to solve the DCP.

The first condition (FC1) occurs if a parameter identification step is attempted when no parameter exists such that the model fits the data. This means that the data-model pair (y, M) under consideration is not in the domain, Γ, of the operator ID. For example, in the first correction step, if $\hat{e}_{2,\text{cal}}$ is such that there is no $\tilde{p}$ that satisfies $\overline{y}_{2,\text{test}} = M_{\text{test}}(\hat{e}_{2,\text{cal}}, \tilde{p})$, then the parameter estimation step fails at this point. In terms of Equation (7), this failure condition occurs if it occurs for *any* point e in $E_{2,\text{cal}}$.

The second failure condition (FC2) occurs if correction step two is able to produce a trajectory not equal to the true trajectory, i.e., $\hat{y}_{1,\text{test}} \neq \overline{y}_{1,\text{test}}$. In terms of the set Y_1 defined in Equation (8), this means that Y_1 contains at least one element that is not equal to $\overline{y}_{1,\text{test}}$. ◇

IV. IDENTIFIABILITY CONDITIONS

In this section, we show that the SGI property is not necessary for the calibration-correction method to solve the DCP. This will be stated as a corollary of the main result of this section (Theorem 1), which gives conditions on the sets of parameters obtained during the calibration-correction method such that the method solves the DCP.

The results rely on two insights. First, the set of output indistinguishable parameters is an equivalence class with respect to the problem of fitting model output trajectories to data [1]. In terms of implementation, this means that an *arbitrary* point may be picked from this set, and the model output will fit the data trajectories at this point. Second, the calibration-correction method involves only fitting and predicting the output trajectories, and not the full state trajectories. This allows us to consider the possibility of treating the sets of parameters obtained during the calibration step and the first correction step as equivalence classes with respect to the prediction step (correction step two) of the method. Indeed, we derive conditions under which we may pick arbitrary points from the sets $E_{2,\text{cal}}$, $C_{2,\text{test}}$ and $E_{1,\text{cal}}$ and still have the method solve the DCP.

Theorem 1 (Parameter consistency)**.** *Consider the DCP in the model universe. Furthermore, consider the calibration-correction method, and the sets* $\tilde{\Theta}_{\text{cal}}$, $E_{1,\text{cal}}$, $E_{2,\text{cal}}$ *and* $P'_{2,\text{test}}$. *Define* $\Theta_{i,\text{test}} \triangleq \text{ID}_\theta\left(\overline{y}_{i,\text{test}}, \overline{M}_{\text{test}}(\theta)\right)$ *for* $i = 1, 2$. *Then, the conditions*

$$\tilde{\Theta}_{\text{cal}} \neq \emptyset, \tag{9}$$

$$E_{2,\text{cal}} \subseteq \text{proj}_e\, \Theta_{2,\text{test}}, \tag{10}$$

$$E_{1,\text{cal}} \times P'_{2,\text{test}} \subseteq \Theta_{1,\text{test}}, \tag{11}$$

are necessary and sufficient for the calibration-correction method to solve the DCP.

Proof. We note that solving the DCP using the calibration-correction method simply involves avoiding the failure conditions FC1 and FC2 described in Remark 7. Avoiding FC1 wherever it may occur ensures that the method can be implemented in the first place, and avoiding FC2 means that the method returns only the desired result. Thus, we must show that the conditions (9-11) are necessary and sufficient for avoiding FC1 and FC2.

The necessity of condition (9) follows from the fact that if $\tilde{\Theta}_{\text{cal}} = \emptyset$, then there does not exist a vector (e_1, e_2, p) such that $\overline{y}_{i,\text{cal}} = M_{\text{cal}}(e_i, p)$ for $i = 1, 2$, leading to FC1 at the calibration step. We note in passing that in the model universe, condition (9) always holds.

Next, we prove the necessity of $E_{2,\text{cal}} \subseteq E_{2,\text{test}} \triangleq \text{proj}_e\, \Theta_{2,\text{test}}$. Assume that there exists an $\tilde{e} \in E_{2,\text{cal}}$ such that $\tilde{e} \notin E_{2,\text{test}}$. Thus, there does not exist a $\tilde{p}$ such that $M_{\text{test}}((\tilde{e}, \tilde{p})) = \overline{y}_{2,\text{test}}$. Since the operator $\text{ID}_{p|e=\tilde{e}}$ is only defined on the set $\{(y, M) \mid \exists p : M((\tilde{e}, p)) = y\}$, we see that the map $\text{ID}_{p|e=\tilde{e}}(\overline{y}_{2,\text{test}}, M_{\text{test}}(e, p))$ is not well defined, leading to FC1 at the first correction step.

We prove the necessity of condition (11) as follows. Assume that there exists a $(\tilde{e}, \tilde{p}) \in E_{1,\text{cal}} \times P'_{2,\text{test}}$ such that $(\tilde{e}, \tilde{p}) \notin \Theta_{1,\text{test}}$. Since we use points $\hat{e} \in E_{1,\text{cal}}$ and $\hat{p} \in P'_{2,\text{test}}$ to generate the prediction $\hat{y}_{1,\text{test}}$ in the second correction step, it is possible that $\hat{e} = \tilde{e}$ and $\hat{p} = \tilde{p}$. Since $\Theta_{1,\text{test}}$ is the set of all points (e, p) that give the correct trajectory $\overline{y}_{1,\text{test}}$, we have the possibility that $\hat{y}_{1,\text{test}} \neq \overline{y}_{1,\text{test}}$, giving us FC2.

Finally, sufficiency is a simple consequence of the fact that conditions (9-11) address both the points in the method where FC1 could be met, and the point in the method where FC2 could occur. Explicitly, condition (9) allows the calibration step to avoid FC1, condition (10) allows correction step one to avoid FC1, since it implies that for all $\tilde{e} \in E_{2,\text{cal}}$, there exists a $\tilde{p}$ such that $(\tilde{e}, \tilde{p}) \in \Theta_{2,\text{test}}$. Condition (11) enables correction step two to avoid FC2, since it implies that for all $\tilde{e} \in E_{1,\text{cal}}$ and for all $\tilde{p} \in P'_{2,\text{test}}$ we have that $\overline{y}_{1,\text{test}} = M_{\text{test}}(\tilde{e}, \tilde{p})$, implying that the set of all possible predicted trajectories only has the correct trajectory in it, $Y_1 = \{\overline{y}_{1,\text{test}}\}$. □

Remark 8. We can give some physical interpretations of the conditions (9-11). To do this, we first note that condition (11) implies (see the lemma in Appendix 2B of [6])

$$E_{1,\text{cal}} \subseteq \text{proj}_e\, \Theta_{1,\text{test}}, \tag{12}$$

$$P'_{2,\text{test}} \subseteq P'_{1,\text{test}}, \tag{13}$$

where $P'_{1,\text{test}}$ is defined in a similar way to $P'_{2,\text{test}}$.

Condition (9) and (12) may be interpreted to mean that the calibration experiments must be more informative about the ESPs than the test process experiments. This follows from the fact that the sets of output-indistinguishable ESPs obtained from the calibration step are subsets of the corresponding sets from the test processes, $\text{proj}_e\, \Theta_{i,\text{test}}$.

Condition (13) says that the PSP sets for the test process, if estimated by first fixing the ESPs to values obtained at the calibration stage, must agree. Agreement here is defined to be unidirectional, with one set being a subset of another. This is only because the correction being performed is from the candidate environment to the refer-

ence environment. If bidirectional correction (Corollary 2 below) were required, then we would have equality in condition (13).

Finally, condition (11) says that the ESP and PSP coordinates in the set $\Theta_{1,\text{test}}$ can only *covary* outside $E_{1,\text{cal}} \times P'_{2,\text{test}}$, i.e., all the points within this set must belong to $\Theta_{1,\text{test}}$. Covariation is defined in Section V. ◇

Next, we state a few corollaries of the theorem.

Corollary 1 (SGI Sufficiency)**.** *SGI models are sufficient for the calibration-correction method to solve the DCP in the model universe.*

Proof. Since the models are SGI, the nominal model universe parameters uniquely fit the model to the data, and therefore the sets in conditions (9-11) only have single entries. Therefore, these conditions are trivially satisfied:

$$\begin{aligned}
&\tilde{\Theta}_{\text{cal}} = \{(\bar{e}_1, \bar{e}_2, \bar{p}_{\text{cal}})\} \neq \emptyset,\\
&E_{2,\text{cal}} = \{\bar{e}_2\} \subseteq \text{proj}_e\{(\bar{e}_2, \bar{p}_{\text{test}})\} = \text{proj}_e\, \Theta_{2,\text{test}},\\
&E_{1,\text{cal}} \times P'_{2,\text{test}} = \{\bar{e}_1\} \times \{\bar{p}_{\text{test}}\} \subseteq \{(\bar{e}_1, \bar{p}_{\text{test}})\} = \Theta_{1,\text{test}}.
\end{aligned}$$

□

Corollary 2 (Bidirectional Correction)**.** *To be able to correct the test data from either environment to the other requires that:*

$$\begin{aligned}
\tilde{\Theta}_{\text{cal}} &\neq \emptyset,\\
E_{i,\text{cal}} &\subseteq \text{proj}_e\, \Theta_{i,\text{test}}, \qquad i = 1, 2,\\
E_{1,\text{cal}} \times P'_{2,\text{test}} &\subseteq \Theta_{1,\text{test}},\\
E_{2,\text{cal}} \times P'_{1,\text{test}} &\subseteq \Theta_{2,\text{test}}.
\end{aligned}$$

Proof. The proof is a simple union of the sets of conditions implied by Theorem 1 for each direction of correction. □

Remark 9. We note that the condition $P'_{2,\text{test}} \subseteq P'_{1,\text{test}}$ discussed in Remark 8 gets transformed into $P'_{2,\text{test}} = P'_{1,\text{test}}$. ◇

Next we discuss the case of correcting the calibration data itself. This will be important in the next section when we examine the effect of a phenomenon called parameter covariation on the calibration-correction method. There, we will prove that a modified version of the method is able to solve the problem at least for this case, even in the presence of parameter covariation.

Corollary 3 ('Test = Calib' Case)**.** *Consider the DCP for the case where* $\overline{y}_{i,\text{test}} = \overline{y}_{i,\text{cal}}$ *and* $\overline{M}_{\text{test}} = \overline{M}_{\text{cal}}$ *for* $i = 1, 2$. *Furthermore, define* $\Theta_{i,\text{cal}} \triangleq \text{ID}_\theta\left(\overline{y}_{i,\text{cal}}, M_{\text{cal}}(\theta)\right)$ *for* $i = 1, 2$, *and*

$$P'_{2,\text{cal}} \triangleq \bigcup_{\tilde{e} \in E_{2,\text{cal}}} \text{ID}_p\left(\overline{y}_{2,\text{cal}}, M_{\text{cal}}(\tilde{e}, p)\right). \tag{14}$$

Then, the conditions

$$\tilde{\Theta}_{\text{cal}} \neq \emptyset, \tag{15}$$

$$E_{2,\text{cal}} \subseteq \text{proj}_e\, \Theta_{2,\text{cal}}, \tag{16}$$

$$E_{1,\text{cal}} \times P'_{2,\text{cal}} \subseteq \Theta_{1,\text{cal}}, \tag{17}$$

are necessary and sufficient for the calibration-correction method to solve this problem.

Proof. Simply specialize Theorem 1 to this case. □

V. PARAMETER COVARIATION

In this section, we describe parameter covariation (Figure 3), and show that it causes the calibration-correction method to fail. We then discuss an improvement to the method that addresses this issue. We start by defining a device that will be useful for taking slices of parameter sets.

Definition 7 (Cutting Plane)**.** Consider the space of parameters $\mathbb{R}^q$, the vector $\theta \in \mathbb{R}^q$ partitioned into two sets of coordinates $\theta = (\theta_a, \theta_b) \in \mathbb{R}^{q_a} \times \mathbb{R}^{q_b}$ and the subspaces $A \triangleq \mathbb{R}^{q_a} \times \{0\}$ and $B \triangleq \{0\} \times \mathbb{R}^{q_b}$ corresponding to the θ_a and θ_b coordinates respectively. Let $(\tilde{\theta}_a, 0) \in A$. Then, we denote the *cutting plane* generated by shifting the origin of B to $(\tilde{\theta}_a, 0)$ with the notation $\text{cut}_{\theta_b}(\tilde{\theta}_a)$.

Definition 8 (Parameter Covariation)**.** Consider the space of parameters $\mathbb{R}^q$ and the vector $\theta \in \mathbb{R}^q$ partitioned into two sets of coordinates $\theta = (\theta_a, \theta_b) \in \mathbb{R}^{q_a} \times \mathbb{R}^{q_b}$. Consider some set of parameters $\Theta \subseteq \mathbb{R}^q$. If there exist $\tilde{\theta}_{a1}, \tilde{\theta}_{a2} \in \text{proj}_{\theta_a} \Theta$ such that $\text{proj}_{\theta_b}\left(\Theta \cap \text{cut}_{\theta_b}(\tilde{\theta}_{a1})\right) \neq \text{proj}_{\theta_b}\left(\Theta \cap \text{cut}_{\theta_b}(\tilde{\theta}_{a2})\right)$, then Θ is said to have *parameter covariation* of its θ_b coordinates with respect to its θ_a coordinates.

Remark 10. We will often abbreviate parameter covariation to just covariation, and say that parameter coordinates can *covary*. ◇

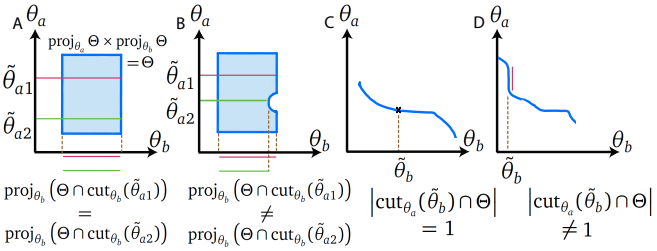

Fig. 3. Parameter covariation. (A) A Cartesian product condition is equivalent to a set not having covariation (Lemma 1). (B) The definition of covariation illustrated. (C) Thin covariation in the θ_a coordinates with respect to the θ_b coordinates. (D) The set in blue does not display thin covariation in the θ_a coordinates with respect to the θ_b coordinates.

Lemma 1. *Let* $\theta = (\theta_a, \theta_b) \in \Theta \subseteq \mathbb{R}^q$ *be a partition of the coordinates of* $\mathbb{R}^q$. *Then, the set* Θ *has covariation of its* θ_b *coordinates with respect to its* θ_a *coordinates if and only if* $\text{proj}_{\theta_a} \Theta \times \text{proj}_{\theta_b} \Theta \neq \Theta$.

Proof. First, we prove the ($\Rightarrow$) direction. Covariation implies that for some $\theta_{a1}, \theta_{a2} \in \text{proj}_{\theta_a} \Theta$ there exists a point $\tilde{\theta}_b \in \text{proj}_{\theta_b} \Theta$ such that

$$\tilde{\theta}_b \in \left(\text{proj}_{\theta_b}\left(\Theta \cap \text{cut}_{\theta_b}(\tilde{\theta}_{a1})\right) \triangle \left(\text{proj}_{\theta_b}\left(\Theta \cap \text{cut}_{\theta_b}(\tilde{\theta}_{a2})\right),\right. \tag{18}$$

where $\triangle$ is the symmetric difference set operation. It further implies that there exists a point $\tilde{\theta}_a \in \{\tilde{\theta}_{a1}, \tilde{\theta}_{a2}\} \subseteq \text{proj}_{\theta_a} \Theta$ such that $(\tilde{\theta}_a, \tilde{\theta}_b) \notin \Theta$. Thus, $\text{proj}_{\theta_a} \Theta \times \text{proj}_{\theta_b} \Theta \neq \Theta$.

Next, we prove the ($\Leftarrow$) direction. Let $(\tilde{\theta}_{a1}, \tilde{\theta}_b) \in \text{proj}_{\theta_a} \Theta \times \text{proj}_{\theta_b} \Theta$ be such that $(\tilde{\theta}_{a1}, \tilde{\theta}_b) \notin \Theta$. Since $\tilde{\theta}_b \in \text{proj}_{\theta_b} \Theta$, there exists a $\tilde{\theta}_{a2} \in \text{proj}_{\theta_a} \Theta$ such that $(\tilde{\theta}_{a2}, \tilde{\theta}_b) \in \Theta$. Thus we have $\tilde{\theta}_b \in \text{proj}_{\theta_b} \left(\Theta \cap \text{cut}_{\theta_b}(\tilde{\theta}_{a2})\right)$ but $\tilde{\theta}_b \notin \text{proj}_{\theta_b} \left(\Theta \cap \text{cut}_{\theta_b}(\tilde{\theta}_{a1})\right)$, which proves the assertion. □

Corollary 4. *The set Θ has covariation of its θ_b coordinates with respect to its θ_a coordinates if and only if it has covariation of its θ_a coordinates with respect to its θ_b coordinates.*

Proof. Using Lemma 1, along with a version of it where the roles of θ_a and θ_b are swapped, leads to this result. □

Remark 11. This equivalence will allow us to refer to sets having covariation with respect to a given partition, such as (e, p). ◇

Next, we show that in the presence of covariation, the calibration-correction method is unable to solve the DCP even in the 'Test = Calib' case of Corollary 3. In particular, we will assume that the restriction of $\Theta_{1,\text{cal}}$ to $E_{1,\text{cal}} \times \text{proj}_p \Theta_{2,\text{cal}}$ has covariation with respect to the (e, p) partition.

Proposition 1. *Consider the 'Test = Calib' case of the DCP. Assume the conditions*

$$\tilde{\Theta}_{\text{cal}} \neq \emptyset, \tag{19}$$

$$P'_{2,\text{cal}} \subseteq \text{proj}_p \Theta_{1,\text{cal}}, \tag{20}$$

$$E_{i,\text{cal}} \subseteq \text{proj}_e \Theta_{i,\text{cal}}, \qquad i = 1, 2, \tag{21}$$

hold, but the set

$$\Theta'_{1,\text{cal}} \triangleq \Theta_{1,\text{cal}} \cap \left(E_{1,\text{cal}} \times \text{proj}_p \Theta_{2,\text{cal}}\right) \tag{22}$$

has covariation with respect to the (e, p) partition. Then, the calibration-correction method fails to solve this problem.

Proof. Condition (20), along with the fact that for the 'Test = Calib' case, $P'_{2,\text{cal}} = \text{proj}_p \Theta_{2,\text{cal}}$, implies that $\text{proj}_p \Theta'_{1,\text{cal}} = P'_{2,\text{cal}}$. Condition (21) implies $\text{proj}_e \Theta'_{1,\text{cal}} = E_{1,\text{cal}}$. Covariation implies that $\text{proj}_e \Theta'_{1,\text{cal}} \times \text{proj}_p \Theta'_{1,\text{cal}} \neq \Theta'_{1,\text{cal}}$. Thus, the proper subset relation $\Theta'_{1,\text{cal}} \subsetneq E_{1,\text{cal}} \times P'_{2,\text{cal}}$ holds, and therefore there exists $(\tilde{e}, \tilde{p}) \in E_{1,\text{cal}} \times P'_{2,\text{cal}}$ such that $(\tilde{e}, \tilde{p}) \notin \Theta'_{1,\text{cal}} \subseteq \Theta_{1,\text{cal}}$. This implies that $E_{1,\text{cal}} \times P'_{2,\text{cal}} \nsubseteq \Theta_{1,\text{cal}}$, which violates condition (17). □

Next, we show that for a specific type of covariation, which we call *thin* covariation, a modified version of the calibration-correction method is able to solve the DCP for the 'Test = Calib' case.

Definition 9 (Thin Covariation)**.** Let $\Theta \subset \mathbb{R}^q$ be a set of parameters and let $(\theta_a, \theta_b) \in \mathbb{R}^q$ be a partition of the coordinates of $\mathbb{R}^q$. If Θ covaries with respect to this partition and if for all $\tilde{\theta}_b \in \text{proj}_{\theta_b} \Theta$, we have $\left|\text{cut}_{\theta_a}(\tilde{\theta}_b) \cap \Theta\right| = 1$, then we say that the covariation of the θ_a coordinates of Θ is thin with respect to the θ_b coordinates.

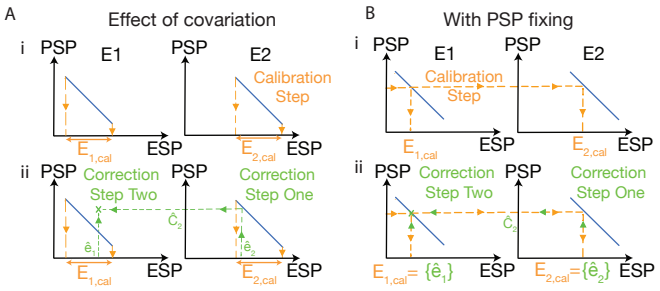

Fig. 4. (A) A schematic description of how thin covariation between the ESP-PSP coordinates in the estimated joint parameter sets can cause calibration-correction to fail at correcting even the calibration data ('Test = Calib' special case). The blue lines in all the plots are the joint ESP-PSP sets of all the parameter values that fit the calibration model to data. (B) How the PSP fixing modification (Definition 10) to the calibration step helps solve this issue. The ESP sets estimated at the calibration step are now generated by first intersecting the parameter sets (blue lines) with a line parallel to the ESP axis ('cutting plane' parallel to the ESP subspace in higher dimensions) centered at an arbitrary PSP value that can be attained, and secondly projecting these intersections to the ESP coordinates for both environments.

Remark 12. We note that if $\Theta \triangleq \text{ID}_\theta(\overline{y}, M(\theta))$, then the condition that for all $\tilde{\theta}_b \in \text{proj}_{\theta_b} \Theta$, we have $\left|\text{cut}_{\theta_a}(\tilde{\theta}_b) \cap \Theta\right| = 1$ is equivalent to the θ_a coordinates of the model $M(\theta_a, \theta_b)$ being SGI for each fixed θ_b. ◇

Remark 12 says that this type of covariation is essentially a statement about the some coordinates being conditionally structurally globally identifiable, despite covarying with respect to the remaining coordinates.

Definition 10 (PSP Fixing)**.** Consider the sets $\Theta_{i,\text{cal}} \triangleq \text{ID}_\theta\left(\overline{y}_{i,\text{cal}}, M_{\text{cal}}(\theta)\right)$, $i = 1, 2$ and let $\tilde{p} \in \text{proj}_p \Theta_{1,\text{cal}} \cap \text{proj}_p \Theta_{2,\text{cal}}$. Then, we define PSP fixing as a modification to the calibration step in which the sets $E_{i,\text{cal}} \triangleq \text{proj}_e\left(\text{cut}_e(\tilde{p}) \cap \Theta_{i,\text{cal}}\right)$ for $i = 1, 2$.

Proposition 2. *Consider the sets $\Theta_{i,\text{cal}} \triangleq \text{ID}_\theta\left(\overline{y}_{i,\text{cal}}, M_{\text{cal}}(\theta)\right)$ for $i = 1, 2$, and the partition $\theta = (e, p)$. Assume that the $\Theta_{i,\text{cal}}$ have thin covariation in their p coordinates with respect to their e coordinates. Then, the calibration-correction method with PSP fixing is able to solve the DCP for the 'Test = Calib' case of Corollary 3.*

Proof. Let $\tilde{p} \in \text{proj}_p \Theta_{1,\text{cal}} \cap \text{proj}_p \Theta_{2,\text{cal}}$ and $\tilde{e}_2 \in E_{2,\text{cal}} \triangleq \text{proj}_e\left(\text{cut}_e(\tilde{p}) \cap \Theta_{2,\text{cal}}\right)$. We note that the sets $\text{proj}_p\left(\text{cut}_p(\tilde{e}_2) \cap \Theta_{2,\text{cal}}\right) = \text{ID}_p(\overline{y}_{2,\text{cal}}, M_{\text{cal}}(\tilde{e}_2, p))$ are equal by definition. Now, pick an arbitrary point $\tilde{p}' \in \text{proj}_p\left(\text{cut}_p(\tilde{e}_2) \cap \Theta_{2,\text{cal}}\right)$. It follows that $\tilde{p}' = \tilde{p}$ from the fact that $\tilde{p} \in \text{proj}_p\left(\text{cut}_p(\tilde{e}_2) \cap \Theta_{2,\text{cal}}\right)$ and that the element in $\left|\text{cut}_{\theta_a}(\tilde{\theta}_b) \cap \Theta\right| = 1$ is unique. Thus, the only possible PSP value that can be returned by the first correction step is $\tilde{p}$.

Next, we look at the second correction step. Pick an arbitrary $\tilde{e}_1 \in E_{1,\text{cal}} \triangleq \text{proj}_e\left(\text{cut}_e(\tilde{p}) \cap \Theta_{1,\text{cal}}\right)$. Since the point $(\tilde{e}_1, \tilde{p}) \in \Theta_{1,\text{cal}}$, we have that $\overline{y}_{1,\text{cal}} = \hat{y}_{1,\text{cal}} \triangleq M(\tilde{e}_1, \tilde{p})$,

and FC2 is avoided. □

Example 4. The final column in Figure 2iii shows the improvement in the data correction when PSP fixing is performed. A model universe study with covariation and PSP fixing visualized in three dimensions can be found in [6]. △

VI. DISCUSSION

Our main goal has been to show that global (or even local) structural identifiability is not an *a priori* necessary condition for models to be useful, and have demonstrated this idea on the problem of transforming or *correcting* model behavior between environments. This data transformation framework is not limited to batch correction in cell extracts. Other examples include correction between cell strains, between *in vitro* and *in vivo* environments, or even between wind tunnels [2].

Next, we discuss some limitations and extensions of this work. The results in this paper are stated for the ideal case where the data are noise free, and some method of getting the set of structurally non-identifiable parameters is available. In practice, both of these conditions are only met approximately: the data usually has some noise, and set identification methods only give approximations to the parameter sets (see [8] for a recent example). We suggest co-opting Markov chain Monte Carlo (MCMC) methods, such as emcee [9], to get the parameter distributions. These distributions, with appropriately small noise model parameters, and noise free model universe data, can give good approximations to the true shape of the sets of identified output-indistinguishable parameters. In general, these sets are approximations because they will have 'smeared out' boundaries, no matter how small the noise model term is made. While the subset relations derived in this paper can still be checked, when the arbitrary points are picked near the boundaries of the sets, some error will appear in the corrected trajectories.

These considerations raise the possibility of extending the set based arguments to a probabilistic setting, where the consistency conditions become probabilistic, and both structural and practical identifiability must be considered.

Another practical consideration related to parameter identification is that, in general, the parameter sets to be estimated may be infinitely large, and must be bounded when estimated. This is a standard practice in the use of MCMC to estimate parameters, and is implemented by setting the prior probability distribution to zero outside some hypercube.

The second set of limitations involves notions of how informative the data is [10], such as persistence of excitation and sufficient richness of inputs and initial conditions. These are not explicitly considered here, precisely because we are concerned with what can be accomplished by models when such conditions are not assumed, and structural identifiability need not hold. This, however, raises the intriguing possibility of generalizing such notions to one that leads to "just enough" structural identifiability, in the sense that with such inputs, initial conditions and data, the non-identifiability satisfies the conditions in Theorem 1 or the condition of thin covariation (Definition 9). The investigation of such a generalization is left as future work, and may begin by first defining data to be sufficiently informative if parameter consistency like conditions hold, and then characterizing the properties of such data.

Another possible generalization of this work follows from the fact that the non-necessity of identifiability presented in this paper can be restated for the more generic problem of the prediction of full system behavior from the behavior of component subsystems. For example, the simplest generalization of Theorem 1 would state a Cartesian product like condition between the subsystem parameter non-indentifiabilities and the full system non-identifiabilities. Similarly, covariation and PSP fixing could be generalized to conditions on which parameters must be made identifiable by model reduction or careful experiment design, and which ones may be allowed to stay non-identifiable.

Finally, these results may be specialized to linear systems, and to specific classes of nonlinear systems, such as bilinear systems [11].

ACKNOWLEDGMENT

The authors would like to thank Samuel Clamons, Wolfgang Halter, William Poole, Anandh Swaminathan and Andras Gyorgy for useful discussions.

REFERENCES

[1] S. Vajda, "Structural equivalence of linear systems and compartmental models," *Mathematical Biosciences*, vol. 55, pp. 39–64, July 1981.

[2] R. S. Swanson and C. L. Gillis, "Wind-Tunnel Calibration and Correction Procedures for Three-Dimensional Models," *NACA Wartime Reports*, vol. L4E31, Oct. 1944.

[3] E. Walter and Y. Lecourtier, "Global approaches to identifiability testing for linear and nonlinear state space models," *Mathematics and Computers in Simulation*, vol. 24, pp. 472–482, Dec. 1982.

[4] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinformatics*, vol. 25, pp. 1923–1929, Aug. 2009.

[5] Z. Z. Sun, C. A. Hayes, J. Shin, F. Caschera, R. M. Murray, and V. Noireaux, "Protocols for Implementing an Escherichia coli Based TX-TL Cell-Free Expression System for Synthetic Biology," *Journal of Visualized Experiments : JoVE*, Sept. 2013.

[6] V. Singhal, *Modeling, Computation, and Characterization to Accelerate the Development of Synthetic Gene Circuits in Cell-Free Extracts*. Ph.d. thesis, California Institute of Technology, 2019.

[7] A. Grinsted, "Gwmcmc: An implementation of the goodman and weare mcmc sampler for matlab." GitHub Repository, March 2015.

[8] Y. Hori and R. M. Murray, "A state-space realization approach to set identification of biochemical kinetic parameters," in *Control Conference (ECC), 2015 European*, pp. 2280–2285, IEEE, 2015.

[9] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, "emcee: The MCMC Hammer," *Publications of the Astronomical Society of the Pacific*, vol. 125, pp. 306–312, Mar. 2013. arXiv: 1202.3665.

[10] S. T. Glad and L. Ljung, "Model structure identifiability and persistence of excitation," in *29th IEEE Conference on Decision and Control*, pp. 3236–3240 vol.6, Dec. 1990.

[11] E. D. Sontag, Y. Wang, and A. Megretski, "Input Classes for Identifiability of Bilinear Systems," *IEEE Transactions on Automatic Control*, vol. 54, pp. 195–207, Feb. 2009.