Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Identifiability of population models via a measure theoretical approach

Steffen Waldherr *, **Shen Zeng** **, **Frank Allgöwer** **

* *Institute for Automation Engineering, Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, Magdeburg, Germany*
email: steffen.waldherr@ovgu.de
** *Institute for Systems Theory and Automatic Control, Universität Stuttgart, Pfaffenwaldring 9, Stuttgart, Germany*
email: {shen.zeng,frank.allgower}@ist.uni-stuttgart.de

Abstract: Heterogeneity in cell populations is a major factor in the dynamics of cellular systems in living tissue or microbial colonies. This heterogeneity needs to be taken into account for the interpretation of experimental observations as well as in the construction of predictive models for cellular systems. A common modelling framework for heterogeneous cell population is by an infinite ensemble of single cell models. The state of a cell population is in this framework modelled by the distribution of the single cell states. In this paper we study under which conditions the population model is identifiable, i.e., we can determine the initial distribution of cell states and parameters from a dynamic output distribution. We derive a necessary condition on the single cell model based on the classical observability results from linear and nonlinear control theory. Our results are illustrated via examples.

Keywords: Estimation; Observability; Heterogenous population models; Probability theory

1. INTRODUCTION

Living cells, even in genetically homogeneous populations, show a significant amount of heterogeneity, which makes a mathematical analysis with an average single cell model less meaningful. A common case is in biochemical signal transduction, where the involved proteins are expressed at different levels within individual cells, thereby leading to differential cellular responses to a common stimulus (Avery, 2006; Spencer et al., 2009). We consider models where heterogeneity is due to initial states and parameters of cells that are distributed heterogeneously among the population. These population models are composed of a dynamic single cell model, which is structurally identical for all cells in the population, and a probability distribution which describes the heterogeneity in the initial conditions and parameters (Fredrickson et al., 1967; Ataai and Shuler, 1985; Ramkrishna, 2000). From biochemical modelling, we typically obtain the underlying single cell model as a non-linear system

$$\dot{z}(t) = f(z(t), \theta), \quad y(t) = h(z(t), \theta), \tag{1}$$

where $z(t) \in \mathbb{R}^n$ is the concentration vector of biochemical species, $\theta \in \mathbb{R}^q$ is the vector of constant parameters of the reaction kinetics and $y(t) \in \mathbb{R}^m$ is a measured output (Hasenauer et al., 2011a).

For simpler notation we transform system (1) into the form

$$\dot{x} = F(x), \quad y = H(x) \tag{2}$$

by introducing the generalized cell state

$$x := (z_1, \dots, z_n, \theta_1, \dots, \theta_q),$$

and the generalized vector field

$$F(x) := (f_1(x), \dots, f_n(x), 0, \dots, 0).$$

A heterogeneous cell population is then modelled by the structural dynamics of the underlying single cells (2) together with the distribution of the generalized initial conditions

$$x_0 = (z_1(0), \dots, z_n(0), \theta_1, \dots, \theta_q). \tag{3}$$

We denote in the following $d := n + q$ the dimension of the generalized cell state space.

In many biologically relevant systems, the number of cells within a population is very high, in the range of ten thousands to millions of cells. This justifies that we can model a population with a probabilistic framework, i.e. we can assume the distribution to be a probability distribution. The state of the cell population at time t is thereby commonly described by a dynamic probability density function

$$p(t, \cdot) : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto p(t, x), \tag{4}$$

where $\int_B p(t, x)\, \mathrm{d}x$ is the probability that a cell has an internal state $x \in B$ at time t. It is a well-known fact that this probability density function is governed by a partial differential equation (PDE) called population balance equation (Ramkrishna, 2000) which is derived as

$$\frac{\partial p}{\partial t}(t, x) + \mathrm{div}_x(F(x)p(t, x)) = 0, \tag{5}$$

together with the specification of an initial density

$$p(0, x) = p_0(x). \tag{6}$$

It is remarkable, that although the single cell model is in general non-linear, the resulting PDE describing the population is always linear, in the sense that the solution operator mapping initial density p_0 to the solution $p(t, \cdot)$ is a linear and bounded operator.

Let us briefly note that a range of simulation methods has been developed to solve the population balance equation, or to compute approximations. With so called individual based population models (IBPM), the given single cell model is simulated for a large number of cells, each with different parameter values, according to the probability distribution of the initial conditions (3) (Ataai and Shuler, 1985; Henson, 2003). The focus of IBPM usually is on the single cell level and on a short time scale, where dynamics of cell division and death are often neglected.

An alternative model formalism, specifically for biotechnological applications, is a density based formalism, which yields a system of partial differential equations, so called cell population balance models (PBM) (Fredrickson et al., 1967; Fredrickson and Mantzaris, 2002; Stamatakis, 2010). Nevertheless it is still in general difficult to solve high dimensional PDEs, and therefore PBM often employ extremely simple single cell models or neglect the single cell dynamics completely by assuming stationarity.

Alongside the modelling, recent work has also considered the problem of parameter estimation in population models. A typical goal is to estimate the distribution for the initial conditions (3) from experimental data in the form of so-called population snapshots (Waldherr et al., 2009; Hasenauer et al., 2011a,b). Population snapshots are direct measurements of an output density function representing a marginalization of the cell density function $p(t, x)$. While the algorithms to estimate the initial density from such measurements are quite efficient, there is currently a lack of theoretical analysis tools. Most importantly, it is unknown under what conditions it is in principle possible to perform such an estimation. While there is a broad literature on identifiability of single cell models (Farina et al., 2006; Raue et al., 2009), we are not aware of any results concerning identifiability of the population models discussed above.

In this paper, we study the relation between identifiability of the underlying single cell model (1) and the identifiability of the population model (5). A plausible conjecture is that identifiability of the single cell model would be a necessary condition for the identifiability of the population model. The main contribution of this paper is to provide a theoretical framework in which this conjecture is proved rigorously. Our proof is based on a measure theoretical approach to population models, which is introduced in the next section, before we present the main results in Section 3.

2. POPULATION MODELS IN THE MEASURE THEORETICAL APPROACH

In this section we introduce our probabilistic framework in which we describe heterogeneous cell populations (Zeng, 2013). We consider a probability space consisting of the sample space $\mathbb{R}^d$, the corresponding Borel algebra $\mathcal{B}$ on $\mathbb{R}^d$ and a probability measure $\mathbb{P}_0$. Additionally we shall always assume that the probability measure $\mathbb{P}_0$ has a probability density $p_0 \in L^1(\mathbb{R}^d, \mathbb{R})$, i.e.

$$\mathbb{P}_0(B) = \int_B p_0 \, \mathrm{d}\mu \qquad \text{for all } B \in \mathcal{B},$$

where μ denotes the Lebesgue measure.

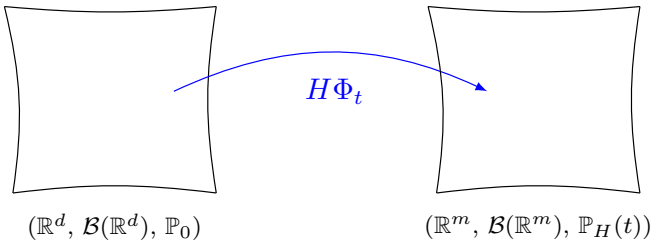

Fig. 1. Illustration of the measure theoretical situation (Zeng, 2013).

As motivated in the introduction, we model heterogeneous cell populations through random initial value problems

$$\begin{aligned} \dot{x} &= F(x), \quad x_0 \sim \mathbb{P}_0, \\ y &= H(x), \end{aligned} \tag{7}$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is the vector field of the generalized single cell dynamics, $\mathbb{P}_0$ is the probability measure of the generalized initial conditions and $H : \mathbb{R}^d \to \mathbb{R}^m$ is the output mapping. Let us also introduce $\Phi_t : \mathbb{R}^d \to \mathbb{R}^d$ as the time-t-mapping associated to the flow of $\dot{x} = F(x)$, i.e.,

$$\Phi_t : x_0 \mapsto x(t; x_0). \tag{8}$$

In typical experiments, such as flow cytometry, we obtain for fixed points t_k in time, measurement outputs $y^{(i)}(t_k)$ for a large number of cells. It is to be stressed that in the measurement process we cannot control which particular cell to measure but for a given instant of time we really can only measure outputs of a large number of cells. This is crucial as it means that there are no single cell trajectories at hand, rendering the problem non-trivial.

Given this type of measurements, however, one can use for example density estimation (Silverman, 1986) to construct an approximation of the probability distribution of the random output $y(t)$, which we shall denote $\mathbb{P}_H(t)$. In other words, we have $y(t) \sim \mathbb{P}_H(t)$. Knowing this output distribution $\mathbb{P}_H(t)$, it is our goal to reconstruct the initial distribution $\mathbb{P}_0$. Let us therefore first discuss the connection between both distributions, cf. (Zeng, 2013).

From probability theory we know that $\mathbb{P}_H(t)$ is the so-called push-forward distribution of the initial distribution $\mathbb{P}_0$ under the composition $H\Phi_t$. This circumstance will be briefly repeated in more detail in the following with the help of Figure 1. There, on the left-hand side we depict the probability space that describes the randomness in the initial conditions x_0. Roughly speaking, the randomness in the initial conditions is propagated to the output $y(t)$, and this is clearly via the mapping $H\Phi_t$. Therefore we have defined on the left probability space the random variable $H\Phi_t(x)$ taking values on a second probability on the right. The randomness in the outputs $y(t)$ is then described by $\mathbb{P}_H(t)$ on the right probability space, which is given as the canonical push-forward distribution as follows

$$\mathbb{P}_H(t)(B_y) := \mathbb{P}_0((H\Phi_t)^{-1}(B_y)), \tag{9}$$

for each $B_y \in \mathcal{B}(\mathbb{R}^m)$. Our measure theoretical framework is now complete and we can proceed to study under which conditions on our measurement function H we can reconstruct the initial distribution from the distributions of the output. This problem is termed identifiability problem.

3. IDENTIFIABILITY OF POPULATION MODELS

3.1 Definition of structural identifiability

For the above setup, several methods have been proposed for estimating the initial probability distribution $\mathbb{P}_0$ from knowledge of $\mathbb{P}_H(t)$ (Waldherr et al., 2009; Hasenauer et al., 2011a,b; Zechner et al., 2012). Even though Zechner et al. (2012) report an identifiability problem found by direct simulation, these estimation methods paid little attention to identifiability properties. In this section we introduce for the first time the concept of structural identifiability of a heterogeneous cell population and derive criteria based on the underlying single cell models. The key to our investigation is the measure theoretical approach introduced in the previous section.

Since the generalized cell state x contains both the species concentrations and the parameters, we can relate identifiability of the single cell model (1) to observability of the extended model (2). We propose a definition of structural identifiability of a heterogeneous population, which is motivated from the notion of indistinguishability in non-linear observability theory in the finite-dimensional case (Zeng, 2013).

Definition 1. (Zeng, 2013). A population model is called *structurally identifiable*, if the following implication holds:

$$(\forall\, t \geq 0 \,:\, \mathbb{P}_H(t;\mathbb{P}_0') = \mathbb{P}_H(t;\mathbb{P}_0'')) \quad \Rightarrow \quad \mathbb{P}_0' = \mathbb{P}_0'', \quad (10)$$

where $\mathbb{P}_0'$ and $\mathbb{P}_0''$ are arbitrary probability distributions.

Thus, if a population model is structurally identifiable, the equality of the output distributions for all times should imply the equality of the underlying initial distributions.

Due to the definition of the output distribution (9), this is equivalent to the statement that the equality

$$\mathbb{P}_0'((H\Phi_t)^{-1}(B_y)) = \mathbb{P}_0''((H\Phi_t)^{-1}(B_y)),$$

for all $t \geq 0$ and all $B_y \in \mathcal{B}(\mathbb{R}^m)$ implies $\mathbb{P}_0' = \mathbb{P}_0''$. Finally, since we assume that the probability measures $\mathbb{P}_0'$ and $\mathbb{P}_0''$ have probability densities, we can rewrite structural identifiability as the following property: If the equality

$$\int_{(H\Phi_t)^{-1}(B_y)} p_0' \,\mathrm{d}\mu \;=\; \int_{(H\Phi_t)^{-1}(B_y)} p_0'' \,\mathrm{d}\mu \qquad (11)$$

holds true for all $t \geq 0$ and for all $B_y \in \mathcal{B}(\mathbb{R}^m)$, then we have equality of the densities p_0' and p_0'' in L^1-sense.

3.2 Structural identifiability with a linear cell model

A necessary condition In this section, for purposes of illustration, we first confine ourselves to linear single cell models, i.e. we assume that $F(x) = Ax$ and $H(x) = Cx$ with $n \times n$-matrix A and $m \times n$-matrix C. As we will see, in this case it is quite straightforward to relate the structural identifiability of a cell population model to the observability of the underlying single cell model.

Theorem 1. (Necessary condition (Zeng, 2013)). Suppose a given cell population model is structurally identifiable, i.e. equality (11) for all $t \geq 0$ and for all $B_y \in \mathcal{B}(\mathbb{R}^m)$ does imply the equality $p_0' = p_0''$. Then (A, C) has to be observable.

Proof. Our proof strategy is to show that under the assumption that (A, C) is not observable, there exist probability densities $p_0' \neq p_0''$ for which equation (11) is true. First we fix an arbitrary probability density p_0'.

It is well-known that the pair (A, C) not being observable is equivalent to the fact that the observability map

$$Ce^{A(\cdot)} : x_0 \mapsto Ce^{A(\cdot)}x_0 = y(\cdot)$$

is not injective, or equivalently (due to linearity), that

$$\ker Ce^{A(\cdot)} = \bigcap_{t\geq 0} \ker Ce^{At} \text{ is non-trivial.}$$

Therefore we can pick a non-zero vector $v \in \bigcap_{t\geq 0} \ker Ce^{At}$, and given that define our second probability density by

$$p_0''(x) := p_0'(x + v).$$

Now we have $p_0' \neq p_0''$, while for all $t \geq 0$ and $B_y \in \mathcal{B}(\mathbb{R}^m)$ we have

$$\begin{aligned}\int_{(Ce^{At})^{-1}(B_y)} p_0''(x) \,\mathrm{d}\mu \;&=\; \int_{(Ce^{At})^{-1}(B_y)} p_0'(x+v) \,\mathrm{d}\mu \\ &= \int_{v+(Ce^{At})^{-1}(B_y)} p_0' \,\mathrm{d}\mu.\end{aligned}$$

Lastly, we observe that

$$v + (Ce^{At})^{-1}(B_y) = (Ce^{At})^{-1}(B_y),$$

since $v \in \ker Ce^{At}$ for all $t \geq 0$. Thus we have shown that (11) is true for all $t \geq 0$ and $B_y \in \mathcal{B}(\mathbb{R}^m)$, while $p_0' \neq p_0''$. □

Example: A simple gene expression model We study the single cell model given by

$$\dot{z} = k - z, \qquad (12)$$

where z corresponds to protein concentration and the constant parameter k is the translation rate, which is assumed to be heterogeneous among cells.

As an illustration we compare measurement of k and measurement of z, which correspond to the output matrices

$$C' = (0\ 1) \qquad \text{and} \qquad C'' = (1\ 0)\,. \qquad (13)$$

We quickly see that (A, C') is not observable, while (A, C'') is observable. As already noted in the proof of Theorem 1, non-observability of (A, C') is equivalent to the fact that the intersection of the kernels $\ker C'e^{At}$ is non-trivial. Due to dimensional reasons we conclude that for all $t \geq 0$ the kernels $\ker C'e^{At}$ are identical. Therefore, for arbitrary $B_y \in \mathcal{B}(\mathbb{R}^m)$, the set $(C'e^{At})^{-1}(B_y)$ consists of a combination of parallel strips as indicated in Figure 2. In this particular case the strips happen to be horizontal.

In conclusion, equality (11) is not strong enough to enforce $p_0' = p_0''$ for the output matrix C'. An initial density can be translated parallel to the x_1-axis without affecting the output densities over time.

Considering the other choice of output, the pair (A, C'') is observable, and the intersection of the kernels $\ker C''e^{At}$ is trivial. We end up in a situation in which by varying t and B_y we can obtain any strip as hinted in Figure 3. From the measurement, we know the integral of the initial density over any of these strips. The identification problem then consists in reconstructing the initial density from the integrals. This is a tomographic problem (Markoe, 2006) and can be handled with tomographic reconstruction methods, as we discuss in a more recent manuscript (Zeng et al., 2014).

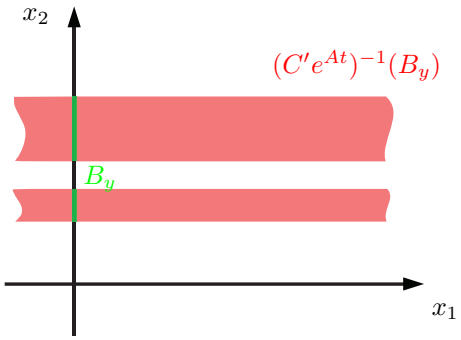

Fig. 2. For given B_y the set $(C'e^{At})^{-1}(B_y)$ is always a combination of parallel strips stretching to infinity (Zeng, 2013).

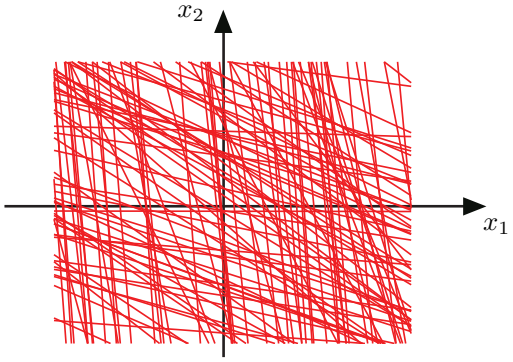

Fig. 3. We can choose $t \geq 0$ and $B_y \in \mathcal{B}(\mathbb{R}^m)$ such that $(C''e^{At})^{-1}(B_y)$ is any strip depicted in this figure (Zeng, 2013). Due to the arrangement of the strips and the fact that the integral of $p_0' - p_0''$ along each strip is zero, we would expect that $p_0' - p_0'' = 0$, i.e. the identifiability of the cell population.

3.3 Structural identifiability with a non-linear cell model

A necessary condition in the non-linear case Let us now turn to the structural identifiability with a non-linear single cell model. The result here is analogous to the linear case: Observability of the single cell model is a necessary condition for the structural identifiability of the population model. In the non-linear case, a significant range of observability definitions exists. Here, we use a definition based on observability codistributions (Hermann and Krener, 1977).

Definition 2. The single cell model (2) is observable, if the observability codistribution

$$Q = \{dH, dL_F H, \ldots, dL_F^{d-1} H\} \tag{14}$$

has dimension equal to d.

It is shown in Isidori (1995) that this condition implies a stratification of the state space, where any two initial conditions from the same stratum are indistinguishable with the output $y = H(x)$, i.e., they yield the same output function over time.

Theorem 2. Suppose that the cell population model (7) is structurally identifiable. Then the single cell model (2) has to be observable.

Proof. Similarly as in the proof of the linear case, we construct a translocation of the probability density for the initial conditions which does not change the output densities. To this end, we first need to transform the model to suitable coordinates.

Non-observability of the single cell model (2) implies that the observability codistribution Q has dimension $d-p < d$. As shown in (Isidori, 1995, Section 1.9), this implies that there exists a (local) coordinate transformation such that the system's dynamics are given by

$$\begin{aligned} \dot{\xi}_1 &= \widetilde{F_1}(\xi_1, \xi_2) \\ \dot{\xi}_2 &= \widetilde{F_2}(\xi_2) \\ y &= \widetilde{H}_2(\xi_2) =: \widetilde{H}(\xi), \end{aligned} \tag{15}$$

with $\xi_1 \in \mathbb{R}^p$ and $\xi_2 \in \mathbb{R}^{d-p}$. We note that any probability measure $\mathbb{P}_0$ as introduced in Section 2 can be represented by a probability density function $\widetilde{p}$ over the space of the transformed coordinates $\xi = (\xi_1, \xi_2)$ and vice versa. Thus, consider an arbitrary density $\widetilde{p}_0'(\xi)$ and let

$$\widetilde{p}_0''(\xi) = \widetilde{p}_0'(\xi + v), \tag{16}$$

with non-zero vector $v \in \operatorname{span}(\{e_1, \ldots, e_p\})$, where e_i denotes the i'th unit vector of $\mathbb{R}^d$. Due to the structure of the system (15), we have

$$(\widetilde{H}\widetilde{\Phi}_t)(\xi) \in \widetilde{B}_y \;\Leftrightarrow\; (\widetilde{H}\widetilde{\Phi}_t)(\xi + v) \in \widetilde{B}_y. \tag{17}$$

This implies that

$$v + (\widetilde{H}\widetilde{\Phi}_t)^{-1}(\widetilde{B}_y) = (\widetilde{H}\widetilde{\Phi}_t)^{-1}(\widetilde{B}_y). \tag{18}$$

Finally we find that

$$\begin{aligned} \int_{(\widetilde{H}\widetilde{\Phi}_t)^{-1}(\widetilde{B}_y)} \widetilde{p}_0''(\xi)\, \mathrm{d}\mu &= \int_{(\widetilde{H}\widetilde{\Phi}_t)^{-1}(\widetilde{B}_y)} \widetilde{p}_0'(\xi + v)\, \mathrm{d}\mu \\ &= \int_{v + (\widetilde{H}\widetilde{\Phi}_t)^{-1}(\widetilde{B}_y)} \widetilde{p}_0'(\xi)\, \mathrm{d}\mu. \end{aligned}$$

Writing this equality in original coordinates, we see that it is nothing but (11). Thus the system is not structurally identifiable. □

Example: A more detailed model of gene expression A more detailed model of gene expression compared to the previous example (12) includes transcription and translation as separate steps, yielding the differential equation

$$\begin{aligned} \dot{z}_1 &= k_1 - z_1 \\ \dot{z}_2 &= k_2 z_1 - z_2, \end{aligned} \tag{19}$$

where z_1 is the mRNA concentration and z_2 the protein concentration. Heterogeneity among individual cells is observed in the transcription rate, represented by the parameter k_1 in the model (19), and has been attributed to differences in the current metabolic state of individual cells (das Neves et al., 2010). A similar level of heterogeneity presumably affects the translation rate, represented by k_2 in the model (19). According to (2), we get the extended single cell model

$$\begin{aligned} \dot{z}_1 &= k_1 - z_1 \\ \dot{z}_2 &= k_2 z_1 - z_2 \\ \dot{k}_1 &= 0 \\ \dot{k}_2 &= 0, \end{aligned} \tag{20}$$

with generalized cell state $x = (z_1, z_2, k_1, k_2)$. Due to the product $k_2 z_1$ of a heterogeneous parameter and an intracellular variable, this model is non-linear.

Let us consider the measurement of the protein concentration as output, defined as

$$y = H(x) = z_2. \tag{21}$$

In order to test for observability of (20) with this output, we construct the observability map q given by

$$(y, \dot{y}, \ddot{y}, \dddot{y}) = q(z_1, z_2, k_1, k_2) \tag{22}$$

and check for its rank (Nijmeijer and van der Schaft, 1990). Thereby, a non-maximal rank of q corresponds to a dimension of the observability codistribution Q smaller than d. As derivatives of y, we obtain

$$\begin{aligned} \dot{y} &= k_2 z_1 - z_2 \\ \ddot{y} &= k_2 k_1 - 2k_2 z_1 + z_2 \\ \dddot{y} &= -2k_2 k_1 + 3k_2 z_1 - z_2. \end{aligned} \tag{23}$$

Thus the observability map q for this model only depends on the three distinct terms z_2, $k_2 k_1$, and $k_2 z_1$, and its rank is at most 3. Thus, the extended system (20) is not observable through the output y, and the corresponding cell population model is not structurally identifiable.

Based on the output derivatives in (23), we note that for the single cell model, two initial conditions x' and x'' which are related by

$$\begin{aligned} z_1' &= \alpha z_1'' \\ k_1' &= \alpha k_1'' \\ k_2' &= \alpha^{-1} k_2'', \end{aligned} \tag{24}$$

with $\alpha \neq 0$, yield the same output trajectory $y(t)$. From this observation, we can construct two different initial densities $p_0'(x)$ and $p_0''(x)$ for the population model, which yield the same output densities $\mathbb{P}_H(t)$ over time.

We propose that any two initial densities $p_0'(x)$ and $p_0''(x)$ which are related by

$$p_0''(z_1, z_2, k_1, k_2) = c p_0'(\alpha z_1, z_2, \alpha k_1, \alpha^{-1} k_2) \tag{25}$$

yield the same output distributions $\mathbb{P}_H(t)$ over time. Thereby, c is a normalization factor to ensure that

$$\int_{\mathbb{R}^4} p_0''(x)\, \mathrm{d}\mu = 1. \tag{26}$$

A simulation of the population model with the nominal and perturbed initial densities has been performed with an approach based on sampling of the initial condition, numerical solution of the single cell model, and subsequent density estimation for the output variable (Hasenauer et al., 2011a). The initial distribution chosen here assigns a probability according to the density functions shown in Figure 4 to the variables k_1, k_2, and $z_1(0)$, and probability 1 to the initial condition $z_2(0) = 0$. We used 1000 samples per initial density for the simulations of this example. The density functions shown in Figure 4 are density estimates from the samples which were used in the simulations. Small ripples in the density functions are an effect of sampling and subsequent density estimation. Figure 5 illustrates that the resulting output densities for the nominal initial density p_0' and the perturbed initial density p_0'' are identical over time. The slight differences between the nominal and perturbed output densities seen in Figure 5 are due to the sampling-based simulation approach.

The example illustrates the close correspondence between observability of the single cell model and identifiability of the population model. Based on an observability analysis of the single cell model (20), we could construct indistinguishable initial densities for the population model even without resorting to the coordinate transformation and

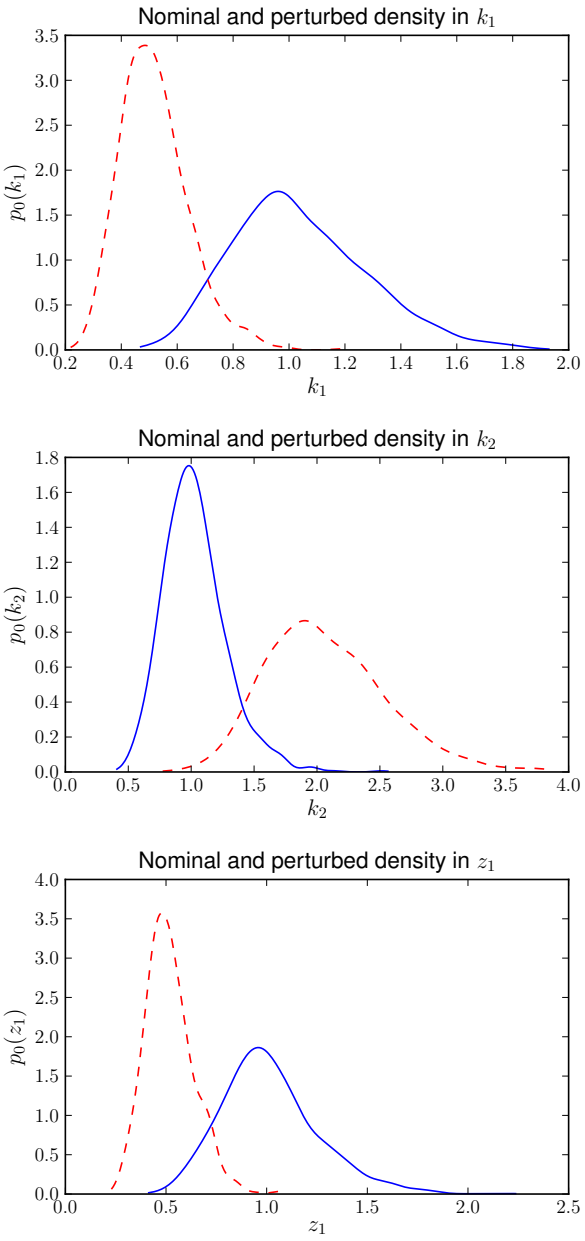

Fig. 4. Initial density functions. Blue full lines show the nominal initial density function p_0', a log-normal distribution centered at the expected values $\mathbb{E}[z_1(0)] = 1$, $\mathbb{E}[k_1] = 1$, and $\mathbb{E}[k_2] = 1$ with a standard deviation of 0.1. Red dashed lines show the perturbed initial density function p_0'', where k_1 and $z_1(0)$ have been scaled by a factor 0.5, and k_2 by a factor 2.

the shifting of the initial density in the transformed coordinates used in the proof of Theorem 2.

4. CONCLUSIONS

We studied the identifiability problem for density-based population balance models. Our study employed a measure theoretical approach, where the state of the population is described within a probability space. With this approach, we proposed a notion of identifiability which is comparable with observability for finite-dimensional systems.

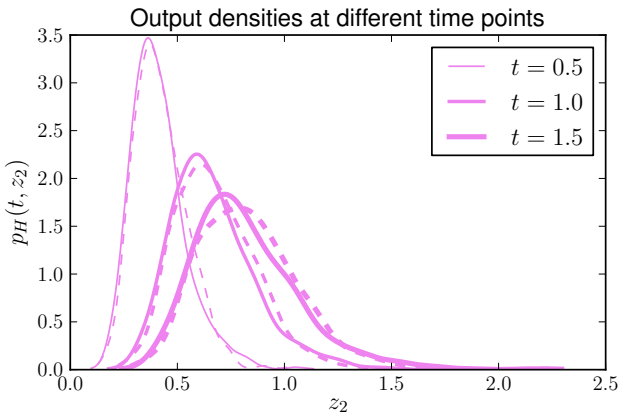

Fig. 5. Output density functions $p_H(t)$ at different points in time. Full lines: simulation from nominal initial density p_0'. Dashed lines: simulation from perturbed initial density p_0''.

Our main result is having shown that identifiability of the single cell model is a necessary condition for identifiability of the population model. Here, identifiability of the single cell model is understood as observability of an extended single cell model with trivial dynamics for the model parameters. The key step in the proof is to show that non-identifiability of the single cell model means that the initial density in the population model can be translated within the unobservable subspace without any effect on the output density over time. With non-linear single cell models, the translation can be done in an appropriately chosen coordinate system for the single cell model. In the example, we have shown that an identifiability analysis of the single cell model yields valuable insights into indistinguishable initial densities even without the formal construction used in the proof.

It is intriguing to speculate that identifiability of the single cell model should also be sufficient for identifiability of the population model. However, a sufficient condition based on tomography theory that we discovered recently (Zeng et al., 2014) only asserts identifiability of the population model for a sufficiently high number of measured variables. That still leaves a gap between the necessary and sufficient conditions to be explored further.

REFERENCES

Ataai, M.M. and Shuler, M.L. (1985). Simulation of CF-STR through development of a mathematical model for anaerobic growth of *Escherichia coli* cell populations. *Biotechnol. Bioengin.*, 27(7), 1051–1055.

Avery, S.V. (2006). Microbial cell individuality and the underlying sources of heterogeneity. *Nat. Rev. Microbiol.*, 4(8), 577–587.

das Neves, R.P., Jones, N.S., Andreu, L., Gupta, R., Enver, T., and Iborra, F.J. (2010). Connecting variability in global transcription rate to mitochondrial variability. *PLoS Biol*, 8(12), e1000560.

Farina, M., Findeisen, R., Bullinger, E., Bittanti, S., Allgöwer, F., and Wellstead, P. (2006). Results towards identifiability properties of biochemical reaction networks. In *Proc. of the 45th IEEE Conf. on Dec. and Control, San Diego, USA*, 2104–2109.

Fredrickson, A.G. and Mantzaris, N.V. (2002). A new set of population balance equations for microbial and cell cultures. *Chem. Eng. Sci.*, 57(12), 2265–2278.

Fredrickson, A., Ramkrishna, D., and Tsuchiya, H. (1967). Statistics and dynamics of procaryotic cell populations. *Math. Biosci.*, 1(3), 327–374.

Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P., and Allgöwer, F. (2011a). Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, 12, 125.

Hasenauer, J., Waldherr, S., Doszczak, M., Scheurich, P., Radde, N., and Allgöwer, F. (2011b). Analysis of heterogeneous cell populations: A density-based modeling and identification framework. *J. Process Control*, 21(10), 1417–1425.

Henson, M.A. (2003). Dynamic modeling of microbial cell populations. *Curr Opin Biotechnol*, 14(5), 460–467.

Hermann, R. and Krener, A. (1977). Nonlinear controllability and observability. *IEEE Trans. Autom. Control*, 22, 728–740.

Isidori, A. (1995). *Nonlinear Control Systems*. Springer-Verlag, London, 3rd edition.

Markoe, A. (2006). Analytic tomography. In *Encyclopedia of Mathematics and its Applications*, volume 106. Cambridge University Press, Cambridge.

Nijmeijer, H. and van der Schaft, A.J. (1990). *Nonlinear dynamical control systems*. Springer- Verlag, New York.

Ramkrishna, D. (2000). *Population Balances: Theory and Applications to Particulate Systems in Engineering*. Academic Press, San Diego, USA.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinf.*, 25(25), 1923–1929.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Spencer, S.L., Gaudet, S., Albeck, J.G., Burke, J.M., and Sorger, P.K. (2009). Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, 459(7245), 428–432.

Stamatakis, M. (2010). Cell population balance, ensemble and continuum modeling frameworks: Conditional equivalence and hybrid approaches. *Chem. Eng. Sci.*, 65(2), 1008–1015.

Waldherr, S., Hasenauer, J., and Allgöwer, F. (2009). Estimation of biochemical network parameter distributions in cell populations. In *Proc. of the 15th IFAC Symp. Syst. Ident. (SYSID)*, 1265–1270. Saint-Malo, France.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., and Koeppl, H. (2012). Moment-based inference predicts bimodality in transient gene expression. *Proc Natl Acad Sci U S A*, 109(21), 8340–8345.

Zeng, S., Waldherr, S., and Allgöwer, F. (2014). An inverse problem of tomographic type in population dynamics. Submitted.

Zeng, S. (2013). *Identifiability and sensitivity analysis of heterogeneous cell population models*. Master's thesis, University of Stuttgart. Supervised by S. Waldherr and F. Allgöwer.