Inference of Kinetics in Population Balance Models using Gaussian Process Regression ⋆

Michiel Busschaert *, **Steffen Waldherr** *,**

* *Department of Chemical Engineering, KU Leuven, 3001 Leuven, Belgium (e-mail: {michiel.busschaert, steffen.waldherr}@kuleuven.be)*
** *Division of Molecular Systems Biology, University of Vienna, 1030 Vienna, Austria (e-mail: steffen.waldherr@univie.ac.at)*

Abstract: Population balance models are used to describe systems composed of individual entities dispersed in a continuous phase. Identification of system dynamics is an essential yet difficult step in the modeling of population systems. In this paper, Gaussian processes are utilized to infer kinetics of a population model, including interaction with a continuous phase, from measurements via non-parametric regression. Under a few conditions, it is shown that the population kinetics in the process model can be estimated from the moment dynamics, rather than the entire population distribution. The method is illustrated with a numerical case study regarding crystallization, in order to infer growth and nucleation rates from varying noise-induced simulation data.

Keywords: Population balance modeling, Gaussian process regression, Crystallization, Systems identification, Moment dynamics

1. INTRODUCTION

Systems consisting of multiple discontinuous entities are generally more difficult to mathematically model and simulate in comparison to strictly continuous systems. Population balance modeling is a well-known method to predict the distribution of the discontinuous entities along individual member properties (Ramkrishna (2000)). This modeling approach has been applied in a wide variety of fields, and in particular, in chemical engineering. In this field, the population balance model is often part of an overall process model. The intrinsic physical nature of the system as well as the process conditions are described by a set of parameters. Those parameter may subsequently be used in the process model for simulation and control. This application of population balance models is known as the solution to the *forward problem*. A different application, known as the *inverse problem*, is in the case when population measurements are available (from the physical process), from which one tries to estimate population dynamics included in the model. To this end, the inverse problem is highly relevant in systems identification. Unfortunately, the inverse problem has proven to be ill-posed in several applications (Ramkrishna (2000); Kostoglou and Karabelas (2005)).

One particular chemical system where population balance models are often studied, is in crystallization. In this process, a dissolved compound aggregates to form a dispersed system of crystals. Crystallization is initiated by a change in temperature and/or medium composition via addition of anti-solvent. This process can be applied in either a batch-wise or a continuous setting. The industrially relevant goal of the crystallization process is either to produce crystals (e.g. in pharmaceutical industry), or to remove solute from the medium as separation process (Myerson (2002)). The intrinsic crystallization dynamics can be divided into nucleation and growth. The former describes how new crystals are created from solute, whereas the latter describes how quickly existing crystals grow in size. Solving the inverse problem hereto typically relies on proposing a certain mathematical expression and estimating parameters from measurements, see for example Bari and Pandit (2018) or Savvopoulos et al. (2019). In addition to growth and nucleation, the population is also affected by phenomena taking place between crystals, such as breakage or aggregation.

In this paper, a method is introduced to estimate population dynamics in interaction with a continuous phase from noisy measurement data. To this goal, if a few conditions are satisfied, then this permits a simplified moment-based approach. The method is illustrated in a non-steady-state crystallization example, in which the continuous phase variable corresponds to dissolved solute. The nucleation and growth rates are inferred from numerically simulated measurement data using Gaussian process regression. This regression method, which has its roots in machine learning, allows for non-parametric inference which can be adapted to varying process conditions. Since no prior expression needs to be imposed, the non-parametric approach has a wider and more adaptive application range.

The paper is structured as follows. Section 2 covers the theoretical background required for the method presented in this paper. The first part of this section covers the basics

⋆ This work has been funded by Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO), grant number G066621N.

of population balance models, and states the conditions under which a moment-based approach may be taken. The second part includes a mathematical formulation of Gaussian process regression. In Section 3, a numerical case study is discussed to illustrate the inference of crystallization dynamics using Gaussian process regression in combination with the reduced moment-based process model. Finally, Section 4 summarizes the paper.

2. METHODS

In this section, a brief theoretical background on two relevant concepts is presented. First, the modeling of particulate systems in combination with continuous phase dynamics is introduced. To this goal, so-called population balance equations are used, following the work of Ramkrishna (2000). Second, the concept of Gaussian process regression is briefly covered, primarily based upon the work of Rasmussen and Williams (2006). The notion of linear operators in Gaussian process is also explained.

Notation—Following mathematical notation is used to distinguish between scalar, vector, and matrix variables. Scalars are denoted in lowercase, e.g. x. Vectors are written in lowercase and bold font weighting, e.g., $\mathbf{x}$. Matrices are written in uppercase and bold font weight, e.g. $\mathbf{X}$.

2.1 Population balance modeling

Population balance equations—A particulate system is characterized as any system that consists of dispersed particles which may interact with other particles and a surrounding continuous medium. Such systems are often mathematically represented via population balance models, which arise in different fields with a wide variety of population types, see for example Ramkrishna and Singh (2014) for a selection of some applications.

In population balance models, each individual particle (population member) has an associated particle state, in this text denoted by an n-dimensional state vector $\mathbf{x}(t) \in \Omega^0_{\mathbf{x}} \subset \mathbb{R}^n$. The dynamics of individual particles are described by a function $\mathbf{f}_{\mathbf{x}}$, mathematically expressed as

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}_{\mathbf{x}}(\mathbf{x}, \mathbf{y}, t). \tag{1}$$

The expected number of members in the total population in particle state x is described by the number density function (NDF) $n(t, \mathbf{x}) \in \mathbb{R}^+$. The expected number of particles with state $\mathbf{x}$ belonging to a region $\Omega_{\mathbf{x}}$ is subsequently given by $\int_{\Omega_{\mathbf{x}}} n(t, \boldsymbol{\xi}) d\boldsymbol{\xi}$. In addition to the particle state, the system dynamics may be influenced by the state of one or more continuous phase variables, denoted by an m-dimensional vector $\mathbf{y}(t) \in \mathbb{R}^m$. From conservation principles, a population balance equation (PBE) in combination with its particle state dynamics, can be formulated generally as

$$\frac{\partial n(t, \mathbf{x})}{\partial t} + \nabla_{\mathbf{x}} \cdot (\mathbf{f}_{\mathbf{x}}(\mathbf{x}, \mathbf{y}, t) n(t, \mathbf{x})) = h(\mathbf{x}, \mathbf{y}, n, t). \tag{2}$$

The first term on the left-hand side of the partial differential equation above represents the accumulation over time, the second term represents the change following internal particle dynamics. The right-hand side, $h(\mathbf{x}, \mathbf{y}, n, t)$, indicates the sum over all source/sink processes. This term also includes birth (or death) processes; events in which particles of one state are generated from (or vanish into) particles of different states. Following boundary fluxes are imposed on the NDF:

$$\begin{gathered} -(\mathbf{f}_{\mathbf{x}}(\mathbf{x}, \mathbf{y}, t) \cdot \mathbf{n}_{\mathbf{x}}) n(t, \mathbf{x}) = \dot{n}_0(\mathbf{x}, \mathbf{y}, t), \text{ for } \mathbf{x} \in \partial\Omega^0_{\mathbf{x}}, \\ \mathbf{f}_{\mathbf{x}}(\mathbf{x}, \mathbf{y}, t) n(t, \mathbf{0}) \to \mathbf{0}, \text{ for } \|\mathbf{x}\| \to \infty. \end{gathered} \tag{3}$$

In equation above, $\partial\Omega^0_{\mathbf{x}}$ denotes the boundary region of the domain $\Omega^0_{\mathbf{x}}$, and $\mathbf{n}_{\mathbf{x}}$ is the normal vector at a $\mathbf{x} \in \partial\Omega^0_{\mathbf{x}}$ oriented outside the domain $\Omega^0_{\mathbf{x}}$, see Ramkrishna (2000) for further details. A general formulation of these dynamics is given by

$$\frac{d\mathbf{y}(t)}{dt} = -\int_{\mathbf{x}} \boldsymbol{\alpha}(\mathbf{x}, \mathbf{y}, t) n(t, \mathbf{x}) d\mathbf{x} + \boldsymbol{\gamma}(\mathbf{y}, t). \tag{4}$$

The first term on the right-hand side indicates the sum of interactions with the population via an exchange rate $\boldsymbol{\alpha}$, and the second term on the right-hand side equals a source/sink term, which is irrespective of the population dynamics.

Moment dynamics—Consider the monovariate case in which the particle state $x \in \mathbb{R}^+$ is a scalar. Then, one may define the k^{th} moment $\mu_k(t)$ as

$$\mu_k(t) = \int_0^\infty x^k n(t, x) dx \tag{5}$$

for $k \in \mathbb{N}$. In the case that the particle dynamics f_x are independent of x, it is possible to explicitly derive moment equations from (2) as

$$\begin{aligned} \frac{d\mu_0(t)}{dt} &= \dot{n}_0(y, t) + \int_0^\infty h(x, \mathbf{y}, n, t) dx, \\ \frac{d\mu_k(t)}{dt} &= k f_x(y, t) \mu_{k-1}(t) + \int_0^\infty x^k h(x, \mathbf{y}, n, t) dx. \end{aligned} \tag{6}$$

for $k > 0$. In the case that the source/sink terms are independent of x and linear in $n(t, x)$, then the integral terms in (6) are functions of $\mu_k(t)$, such that the right-hand side of the k^{th} moment differential equations does not depend on higher order moments.

In addition to the assumptions defined above, resulting in the moment equations, a specific expression of the continuous dynamics renders the continuous dynamics closed under the k^{th} moment. This is the case on the condition that the factor $\boldsymbol{\alpha}(x, \mathbf{y}, t)$ in (4) is strictly polynomial in x. Then, (4) and (6) form a closed system of equations for all moments $\mu_k(t)$ where k is less than or equal to the polynomial order of $\boldsymbol{\alpha}(x, \mathbf{y}, t)$ in x. Initial conditions for the moment differential equations are obtained via integration of (5) evaluated with the NDF initial condition.

Example in crystallization—To illustrate the moment dynamics as described above, the approach is presented for a mixed suspension mixed product removal (MSMPR) crystallization model. This example is utilized as case study in Section 3. Further theoretical background on the mathematical modeling of crystallization processes can be found in, for example, Porru and Özkan (2017) or Savvopoulos et al. (2019).

In this MSMPR process, the population consists of crystals which differ in crystal size $x \in [0, \infty)$. The state dynamics only depend on the solute concentration C in the continuous phase. That is, the growth rate G solely depends on C, i.e., $\dot{x} = G(C)$. Technically speaking, the growth rate is

typically a function of the supersaturation which is defined as the ratio of concentration C to C_{sat}. The saturation concentration C_{sat} is a function of temperature and anti-solvent content in the medium, which are both assumed constant in the process operation. The PBE along with the boundary condition at $x = 0$ is given as

$$\begin{aligned} \frac{\partial n(t,x)}{\partial t} + G(C)\frac{\partial n(t,x)}{\partial x} &= -dn(t,x), \\ G(C)n(t,0) &= B(C). \end{aligned} \tag{7}$$

Here, d represents a constant dilution rate (the fraction of tank volume being replenished with new medium per unit of time). $B(C)$ represents a (primary) nucleation rate, this is, the rate at which new crystals are formed out of solute concentration. Secondary nucleation is assumed to be negligible for this application.

The concentration dynamics are defined by the continuous inflow and outflow of solute, and the solute consumed during crystal growth, resulting in

$$\frac{dC(t)}{dt} = -3k_v\rho_C \int_0^\infty G(C)x^2 n(t,x)dx + d(C_{\text{in}} - C(t)). \tag{8}$$

In equation above, k_v is a shape factor (equal to 1 in case of spherical crystals), ρ_C the material density of crystals. C_{in} represents the concentration of the continuous inflow into the tank.

One can verify that the system described by (7) and (8) satisfies the conditions described earlier. This allows to formulate a moment-based reduction of the system, resulting in

$$\begin{aligned} \frac{dC(t)}{dt} &= -3k_v\rho_C G(C)\mu_2(t) + d(C_{\text{in}} - C(t)), \\ \frac{d\mu_0(t)}{dt} &= -d\mu_0(t) + B(C), \\ \frac{d\mu_1(t)}{dt} &= -d\mu_1(t) + G(C)\mu_0(t), \\ \frac{d\mu_2(t)}{dt} &= -d\mu_2(t) + 2G(C)\mu_1(t). \end{aligned} \tag{9}$$

The advantage of this approach, is that a concentration profile can be derived without solving the PBE simultaneously. Furthermore, if one is only interested in the moments of the NDF, the PBE does not need to be solved. In fact, using the method of characteristics (see Appendix A), the NDF can be obtained without numerically solving the PBE directly.

2.2 Gaussian process regression

Gaussian processes—A Gaussian process is a stochastic process where each finite subset of random variables is jointly Gaussian distributed. It can be regarded as a generalization of the finite dimensional Gaussian probability distribution over random vectors to an infinite dimensional vector, which essentially is a function. The concept of probability distribution over a random function can be exploited for non-parametric regression, as shown below. In this text, the theory behind Gaussian process regression is summarized following the function-space view as given by Rasmussen and Williams (2006).

Assume an unknown function $f(\mathbf{x})$ defined over its (multi-dimensional) domain. Then, similar to a Gaussian distribution, a Gaussian process is fully defined by a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$, where both $\mathbf{x}$ and $\mathbf{x}'$ are within the domain of $f(\mathbf{x})$. Then a Gaussian process is denoted as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{10}$$

with $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ defined as

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]. \end{aligned} \tag{11}$$

Assume N measurements $\mathbf{y}_*$ of $f(\mathbf{x})$ are taken for inputs within the set X_*. In this case, the measurements $\mathbf{y}_*$ are estimates of $f(X_*)$ under white noise. Following this, the Gaussian process can be written as a finite joint distribution between the unknown function and measurements,

$$\begin{aligned} \begin{pmatrix} f(X) \\ \mathbf{y}_* \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} \mathbf{m}(X) \\ \mathbf{m}(X_*) \end{pmatrix}, \begin{pmatrix} \mathbf{K}(X,X) & \mathbf{K}(X,X_*) \\ \mathbf{K}(X_*,X) & \mathbf{K}(X_*,X_*) + \sigma_n^2\mathbb{I} \end{pmatrix} \right) \\ &\sim \mathcal{N}\left(\begin{pmatrix} \mathbf{m}(X) \\ \mathbf{m}(X_*) \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^\intercal & \mathbf{K}_{**} + \sigma_n^2\mathbb{I} \end{pmatrix} \right). \end{aligned} \tag{12}$$

Here, X represents a subset of input values in the domain of $f(\mathbf{x})$ where the Gaussian process is evaluated, while X_* represents the set of all measurement inputs $\mathbf{x}_{*,i}$. Then, $K(X, X')$ is a matrix where each element $\mathbf{K}(X, X')_{i,j} = k(X_i, X'_j)$. The value σ_n^2 is added to the diagonal elements of $\mathbf{K}_{**}$ to compensate for measurement noise.

Regression from measured data is achieved via formulating the conditional distribution for $f(X)$ based on the known measurements $\mathbf{y}_*$ at X_*. A property of the Gaussian distribution, is that the conditional remains a Gaussian distribution, with known mean and covariance (see for example Chapter 6 in Deisenroth et al. (2020)), resulting in

$$\begin{aligned} f(X)|\mathbf{y}_*, X_* \sim \mathcal{N}(&\mathbf{m}(X) + \mathbf{K}_*(\mathbf{K}_{**} + \sigma_n^2\mathbb{I})^{-1}(\mathbf{y}_* - \mathbf{m}(X_*)), \\ &\mathbf{K} - \mathbf{K}_*(\mathbf{K}_{**} + \sigma_n^2\mathbb{I})^{-1}\mathbf{K}_*^\intercal). \end{aligned} \tag{13}$$

Covariance kernel—One aspect which has not been discussed so far, is what functions to select for $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$. Direct derivation from their definition in (11) is not possible since $f(\mathbf{x})$ is unknown. In most cases, $m(\mathbf{x})$ can be taken as the zero function. The selection of the covariance function $k(\mathbf{x}, \mathbf{x}')$, also called the kernel of the Gaussian process, determines the properties of the regression result. The kernel $k(\mathbf{x}, \mathbf{x}')$ needs to be chosen such that the covariance matrix in the Gaussian distribution (12) is positive semidefinite. See Rasmussen and Williams (2006) for details on valid kernel selection.

A commonly implemented kernel is the so-called *Squared Exponential* (SE) kernel. This is an isotropic kernel, meaning the covariance is a function only of the distance between $\mathbf{x}$ and $\mathbf{x}'$. In the case of a 1-dimensional input x, the SE kernel is expressed as

$$k(x, x'; \sigma, l) = \sigma^2 \exp\left(-\frac{1}{2}\frac{(x - x')^2}{l^2}\right). \tag{14}$$

Here, σ and l are *hyperparameters* with readily interpretable meaning. The hyperparameter σ^2 indicates an uncertainty variance, while l is a lengthscale that indicates how quickly the covariance decreases the further x' is located from x. The estimated measurement noise variance σ_n^2 can also be regarded as a hyperparameter.

The selection of hyperparameter values follows from an optimization procedure using measurement data, such as cross-validation or maximization of the (logarithm of the) marginal likelihood, see Chapter 5 of Rasmussen and Williams (2006) for further details. Maximization of the log marginal likelihood corresponds to finding the hyperparameters that result in the least uncertainty on the model prediction of the N measurements $\mathbf{y}_*$ in X_*. The log marginal likelihood (with zero mean function $m(x)$) is expressed as

$$\begin{aligned}\log p(\mathbf{y}_*|X_*;l,\sigma,\sigma_n) = &-\frac{1}{2}\mathbf{y}_*^\intercal(\mathbf{K}_{**}+\sigma_n^2\mathbb{I})^{-1}\mathbf{y}_* \\ &-\frac{1}{2}\log|\mathbf{K}_{**}+\sigma_n^2\mathbb{I}| - \frac{N}{2}\log(2\pi).\end{aligned} \tag{15}$$

The gradient of the log marginal likelihood with respect to hyperparameters may be calculated relatively efficiently (Rasmussen and Williams (2006)), such that optimization using a gradient-descent method is beneficial. It is possible though that there are multiple local maxima of (15). As consequence, attention should be paid to possible non-global maxima [1] , which is attempted via locally optimizing for multiple random initial guesses generated using latin hypercube sampling. Local optimization is performed using the `minimize` function from Rasmussen and Williams (2006), which is available online.

Linear operators—An additional advantage of using Gaussian process regression, is that it is possible to regress over multiple functions that relate to each other by a linear operator, see Särkkä (2011). For example, given a function $f(\mathbf{x})$ and any linear operator $\mathcal{L}_\mathbf{x}$ over $\mathbf{x}$, a function $g(\mathbf{x})$ can be defined as

$$g(\mathbf{x}) = \mathcal{L}_\mathbf{x} f(\mathbf{x}). \tag{16}$$

Following from definition (11), the joint Gaussian process between $f(\mathbf{x})$ and $g(\mathbf{x})$ is formulated as

$$\begin{pmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} m(\mathbf{x}) \\ \mathcal{L}_\mathbf{x} m(\mathbf{x}) \end{pmatrix}, \begin{pmatrix} k(\mathbf{x},\mathbf{x}') & k(\mathbf{x},\mathbf{x}')\mathcal{L}_{\mathbf{x}'} \\ \mathcal{L}_\mathbf{x} k(\mathbf{x},\mathbf{x}') & \mathcal{L}_\mathbf{x} k(\mathbf{x},\mathbf{x}')\mathcal{L}_{\mathbf{x}'} \end{pmatrix}\right). \tag{17}$$

In other words, the Gaussian process over functions related via a linear operator $\mathcal{L}_\mathbf{x}$ is determined by applying of the linear operator to the mean and covariance functions. For further regression purposes, (17) can be treated similar to what was discussed earlier.

One particular example — which is implemented within the numerical case study discussed in Section 3 — is where the linear operator corresponding to a derivative w.r.t. x, this is, $\mathcal{L}_x = \frac{d}{dx}$. Then, the derivative $\frac{df(x)}{dx} = g(x)$ may be estimated from measured function values $f(x_*)$ evaluated at x_*. In this case, interpolation of the derivative between the measurements x_* is possible.

3. NUMERICAL CASE STUDY

In this section, the inference of growth and nucleation rate in the setup of a MSMPR crystallizer is presented. In Subsection 3.1, an overview is given on the conditions under which numerical training data are generated induced with artificial noise. In Subsection 3.2, results from growth and nucleation rate inference using artificial training data under varying noise levels are given.

[1] Non-global local maxima correspond with a certain interpretation of the regression (Rasmussen and Williams (2006)), e.g. overfitting.

Table 1. Overview of used process and kinetic parameters, adapted from Savvopoulos et al. (2019) using the conditions of Case 1 and isothermal process operation.

Parameter	Value	Units
k_{G1}	$3.21 \cdot 10^{-4}$	m/s
k_{G2}	$2.58 \cdot 10^{2}$	J/mol
k_{B1}	$1.15 \cdot 10^{-7}$	#/m^3/s
k_{B2}	$7.67 \cdot 10^{4}$	J/mol
k_{B3}	$1.60 \cdot 10^{-1}$	—
T	298.15	K
R	8.314	J/K/mol
C_{sat}	23.375	kg/m^3
d	$7.96 \cdot 10^{-3}$	1/s
k_v	1	—
ρ_C	$1.4 \cdot 10^{3}$	kg/m^3

3.1 Simulation model

To generate training data, the PBE (7) and concentration equation (8) need to be solved over time. Hereto, the system of four ODE's in (9) is solved first, after which the solution $C(t)$ is substituted into the PBE (7) via the method of characteristics (see Appendix A) to determine the solution of the NDF. Although the NDF is not used directly in the kinetics inference, it is used to calculate the population moments. The process is simulated under dynamical conditions based on a step signal in inlet concentration C_{in} at $t = 0$; the initial conditions C_0 and $n_0(x)$ are obtained from the steady-state solutions of (7) and (8) for constant $C_{\text{in}} = 35$ kg/m^3. The dynamical system described by the ODE's (9) is solved using constant input $C_{\text{in}} = 90$ kg/m^3.

The parameters used in the numerical case study are based on the study performed by Savvopoulos et al. (2019), who investigate the crystallization of aspirin in a tubular crystallizer with ultrasound assistance. In that paper, the following expressions are modeled for growth and primary nucleation rate:

$$\begin{aligned} G(C) &= k_{G1}\exp\left(-\frac{k_{G2}}{RT}\right)(C - C_{\text{sat}}), \\ B(C) &= k_{B1}\exp\left(\frac{k_{B2}}{RT}\right)\exp\left(-\frac{k_{B3}}{\log^2(C/C_{\text{sat}})}\right), \end{aligned} \tag{18}$$

with R representing the universal gas constant. Note that nucleation can only occur when $C > C_{\text{sat}}$, whereas the growth rate $G(C)$ is negative when $C < C_{\text{sat}}$, which corresponds with crystals dissolving instead of growing. Parameter values are adopted, from Case 1 in Savvopoulos et al. (2019). See Table 1 for the list of constant parameter values used in this numerical case study. The effect of secondary nucleation has been ignored in this case. The simulated concentration and NDF are sampled every 15 seconds.

Next, noise is intentionally applied to the simulation results to recreate measurements conditions encountered in physical systems. To this end, white zero-mean Gaussian noise is assumed on the concentration measurement. Noise on the population measurement is applied indirectly, since one only requires the zeroth to second moments. Regarding the zeroth moment, Gaussian noise is applied to the exact zeroth moment following from integration of the NDF over the particle size for a fixed time, which relies on the

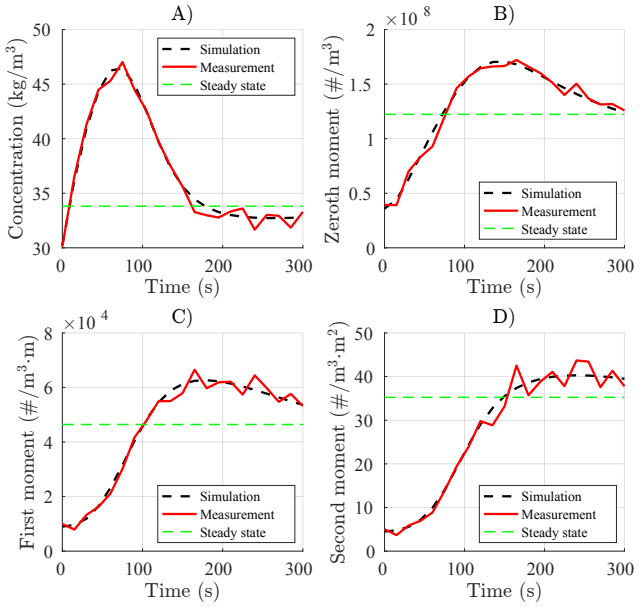


Fig. 1. Measurement example (red solid line) sampled from the simulated process (black dashed line) with following noise parameters: $\sigma_C = 0.5$ kg/m^3, $\sigma_{\mu_0} = 5 \cdot 10^6$#/m^3 and $N_s = 1000$. A) concentration C, B) zeroth moment μ_0, C) first moment μ_1 and D) second moment μ_2.

moment definition from the NDF in (5). Noise on the first and second moment is induced by sampling N_s particle sizes uniformly according to the cumulative NDF. This way, noise is introduced by sampling randomness. The following measurement model is implemented:

$$\begin{aligned} C_{\text{meas}}(t) &= C(t) + \epsilon_C(t) \text{ with } \epsilon_C(t) \sim \mathcal{N}\left(0, \sigma_C^2\right), \\ \mu_{0,\text{meas}}(t) &= \int_0^\infty n(t,x)dx + \epsilon_{\mu_0}(t) \\ &\quad \text{with } \epsilon_{\mu_0}(t) \sim \mathcal{N}\left(0, \sigma_{\mu_0}^2\right), \\ \mu_{1,\text{meas}}(t) &= \frac{1}{N_s}\sum_{i=1}^{N_s} x_i \cdot \mu_{0,\text{meas}}(t), \\ \mu_{2,\text{meas}}(t) &= \frac{1}{N_s}\sum_{i=1}^{N_s} x_i^2 \cdot \mu_{0,\text{meas}}(t). \end{aligned} \tag{19}$$

Figure 1 displays an example of the exact simulation results for concentration and zeroth to second moments, along with a simulated measurement for a given set of noise parameters. Note that the system operates in transient conditions.

3.2 Growth and nucleation rate inference

The general outline of crystallization kinetics inference starts as follows. The growth rate $G(C)$ and nucleation rate $B(C)$ are assumed as separate zero-mean Gaussian processes given by

$$\begin{aligned} G(C) &\sim \mathcal{GP}\left(0, k_G(C,C'); \sigma_G, l_G, \sigma_{n,G}\right), \\ B(C) &\sim \mathcal{GP}\left(0, k_B(C,C'); \sigma_B, l_B, \sigma_{n,B}\right), \end{aligned} \tag{20}$$

where both covariance kernels $k_G(C,C')$ and $k_B(C,C')$ are SE kernels as in (14).

Next, regression is achieved by defining a conditional distribution from training data. The training data is obtained by transformation of the fundamental equations in (9), which results in following training data equations for $G(C)$,

$$\begin{aligned} y_1(C(t)) &= \frac{d(C_{\text{in}} - C(t)) - \dot{C}(t)}{3k_v\rho_C\mu_2(t)} = G(C), \\ y_2(C(t)) &= \frac{d\mu_1(t) + \dot{\mu}_1(t)}{\mu_0(t)} = G(C), \\ y_3(C(t)) &= \frac{d\mu_2(t) + \dot{\mu}_2(t)}{2\mu_1(t)} = G(C), \end{aligned} \tag{21}$$

and for $B(C)$,

$$y_4(C(t)) = d\mu_0(t) + \dot{\mu}_0(t) = B(C). \tag{22}$$

In equations above, the time derivatives are estimated from Gaussian process regression with linear operators, as described in Subsection 2.2. It is observed that using Gaussian process regression for calculating the derivative results in less approximation error compared to discretization using the central difference method (with the exact derivative as evaluation the right-hand sides of (9) as reference). Training data is generated by evaluating (21) and (22) using noise-induced measurements from (19) sampled at inputs $\mathbf{C}_*$. This training data is used to formulate a conditional Gaussian distribution of the Gaussian processes (20) for growth rate $G(C)$ and nucleation rate $B(C)$ evaluated in C, resulting in

$$\begin{aligned} G(C)|\mathbf{y}_{1,*}, \mathbf{y}_{2,*}, \mathbf{y}_{3,*}, \mathbf{C}_* &\sim \mathcal{N}\left(m_G, K_G\right), \\ B(C)|\mathbf{y}_{4,*}, \mathbf{C}_* &\sim \mathcal{N}\left(m_B, K_B\right), \end{aligned} \tag{23}$$

where the means, m_G and m_B, and covariances, K_G and K_B, in each Gaussian process, calculated as in (13), depend on the concentration C where the kinetics are evaluated.

Table 2. Varying σ_C, with $\sigma_{\mu_0} = 5 \cdot 10^6$ #/m^3 and $N_s = 1000$. Median from 50 runs.

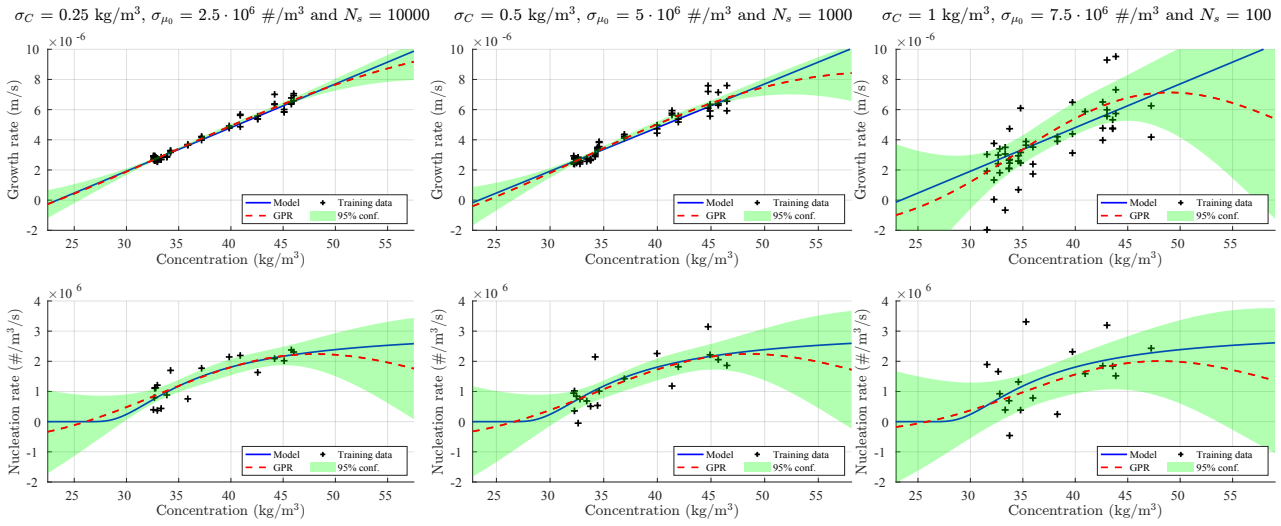
σ_C (kg/m^3)	Growth rate RMSE (10^{-7}m/s)	Nucleation rate RMSE (10^4#/m^3/s)
0	1.59	10.77
0.25	1.47	11.08
0.5	1.71	11.90
1	2.26	14.01
2	3.75	19.74

Table 3. Varying σ_{μ_0}, with $\sigma_C = 0.5$ kg/m^3 and $N_s = 1000$. Median from 50 runs.

σ_{μ_0} (10^6#/m^3)	Growth rate RMSE (10^{-7}m/s)	Nucleation rate RMSE (10^4#/m^3/s)
0	1.28	4.99
2.5	1.36	8.97
5	1.71	11.90
7.5	2.01	16.33
10	2.61	28.64

Table 4. Varying N_s, with $\sigma_C = 0.5$ kg/m^3 and $\sigma_{\mu_0} = 5 \cdot 10^6$ #/m^3. Median from 50 runs.

N_s (—)	Growth rate RMSE (10^{-7}m/s)	Nucleation rate RMSE (10^4#/m^3/s)
100	1.74	12.81
500	1.80	12.36
1000	1.71	11.90
5000	1.78	12.09
10000	1.68	12.24



(a) Low noise (b) Medium noise (c) High noise

Fig. 2. Gaussian process regression (GPR) result for different (relative) levels of noise. Results for both inference of growth rate $G(C)$ inference (top) and nucleation rate $B(C)$ (bottom) are shown. The green band around the regression curve represents a 95% confidence bound from the Gaussian process.

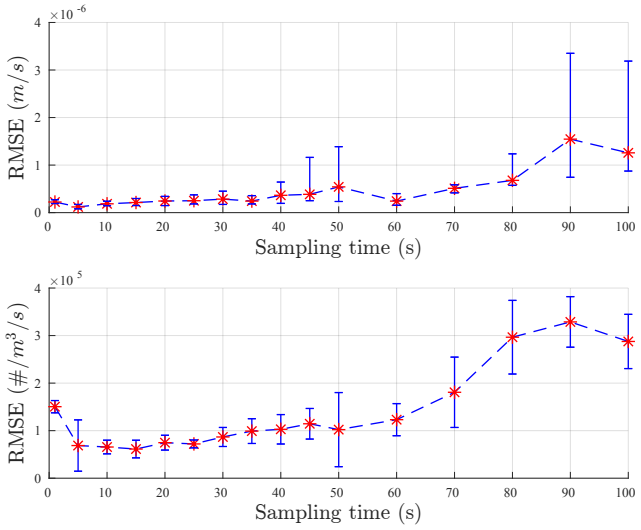


Fig. 3. Median RMSE over 50 runs of regression result for growth rate $G(C)$ (top) and nucleation rate $B(C)$ (bottom) under different sampling times. Lower and upper bounds represent the 25% and 75% quantile, respectively.

The regression is applied to measurements with varying noise levels, this is, for different values of σ_C, σ_{μ_0} and N_s. The performance of the inference results for the growth rate $G(C)$ and nucleation rate $B(C)$, is quantified by the root mean squared error (RMSE), where the mean squared error is obtained via numerical integration of the squared error between the modeled dynamics (18) and the inference result (23), evaluated between the concentration bounds from the noiseless simulation data (in other words, without extrapolating the inference results outside the bounds of the simulated concentration range). Tables 2, 3 and 4 show estimates of the RMSE under varying levels of σ_C, σ_{μ_0} and N_s, respectively, while keeping the two other noise parameters constant. The values are generated by taking the median RMSE's over 50 regressions for different measurement realizations using constant noise parameters. Note that, in some cases, regression seemed to fail, likely due to non-global maxima in hyperparameter optimization (see Subsection 2.2).

As seen from Tables 2 to 4, an increase in σ_C and σ_{μ_0} generally increases the RMSE for both growth and nucleation rate estimate, although in a slightly different manner. Noise on concentration measurement does not only result in inaccuracy on training data y_1, but it affects the covariance kernel input as well. The issue of input noise has been addressed by McHutchon and Rasmussen (2011), who propose a correction by increasing the variance where the output gradient is large. Noise on the zeroth moment from the NDF measurements directly affects the training data y_2 for the nucleation rate inference, and indirectly affects the estimates for the first and second moment used in all training data for the growth rate. Finally, variation on the sampling rate N_s slightly affects growth rate inference, although not as much as variation on the other parameters. For reference, in the case of noiseless data, the RMSE for growth and nucleation rate are respectively $2.25 \cdot 10^{-10}$ m/s and $8.29 \cdot 10^3$ #/m^3/s. Inference in this case is nearly perfect, and the non-zero RMSE's are due to the approximation of the derivatives used in the training data.

For illustration, an example of the regression results is visualized in Figure 2 for different degrees of measurement noise. A higher degree of noise, such as in Figure 2c, results in less accurate inference of both growth and nucleation rate. In case of higher noise levels, the training data is more spread out around the modeled curve compared to cases with lower noise parameters. Note that this phenomena shows up regarding the width of the 95% confidence bands, which is a measure for the optimized hyperparameters σ_G and σ_B.

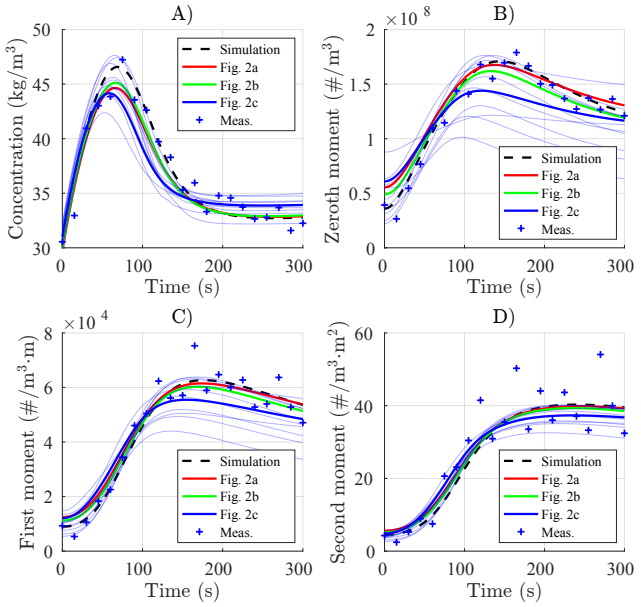


Fig. 4. Model simulation using kinetics from (18) (black dashed line) compared with simulation using regression results (23), with low/medium/high noise levels as in Figure 2 (solid lines). Initial conditions are determined as the steady-state condition with input $C_{\text{in}} = 35$ kg/m^3 using modified kinetics. For illustration, measurement data (blue crosses) and simulation with sampled kinetics (transparant blue lines) under relative high noise levels are shown, a. A) concentration C, B) zeroth moment μ_0, C) first moment μ_1 and D) second moment μ_2.

Naturally, the performance of the inference depends on the time interval at which measurements are sampled. As mentioned, the case study uses a uniform sampling time of 15 seconds. The performance under different sampling times is evaluated and characterized by the RMSE. Figure 3 summarizes RMSE values from 50 runs each visually for uniform sampling times between 1 to 100 seconds. Up to around a sampling time of 40 seconds, the performance is more or less accurate and generally precise. In physical experiments, sampling rates are often limited by the available measurement techniques, this in contrast to numerical experiments. In the former, the sampling time should still be sufficiently low in order to get reasonable estimates for the time derivatives in (21) and (22). The Gaussian process framework allows for training data of different experiments to be combined for inference. In case the possible sampling time is too restrictive, multiple experiments at steady-state conditions could be utilized to generate training data.

In addition, the solution of the concentration and moments equations (9) (and by extension also the NDF) can be determined using the regression results for growth rate $G(C)$ and nucleation rate $B(C)$ instead of the expressions in (18). Using the mean of the Gaussian process regression results under different noise levels shown in Figure 2, the reduced process model is solved, see Figure 4. The inference under low and medium noise levels results in quite similar behaviour to the original simulation. Under higher noise, the simulation is still reasonably well, although less accurate compared to the kinetics from (18). The uncertainty on the inference result (see the 95% confidence bands in Figure 2) is illustrated by simulations with samples from the growth and nucleation rate regression. Samples are drawn jointly according to the conditional Gaussian distribution as in (23). In Figure 4, ten random samples are drawn from the regression under the highest noise level from Figure 2, and used for simulation. Notably, regarding the zeroth moment, a rather broad distribution around the simulation using the mean of the Gaussian process regression is visible.

4. CONCLUSION AND OUTLOOK

In this paper, the concept of using Gaussian process regression for the inverse problem of population balance models is illustrated. In Section 2, a general introduction to population balance models in interaction with a continuous phase is given. It is shown, that under some model assumptions, one can derive a process model based on the moment dynamics. This significantly reduces the problem complexity, since only a few ordinary differential equations are obtained, instead of both partial and ordinary differential equations. Next, an introduction to Gaussian process regression is given, and it is shown how linear operators can be combined in Gaussian processes, which is later used to estimate derivatives from noise-induced measurements. In Section 3, the use of Gaussian process regression on a numerical example related to crystallization of aspirin based on Savvopoulos et al. (2019). It is shown that at all noise levels, a reasonable estimate is mostly obtained, although overall less accurate for higher degrees of noise.

The main advantages of this method are a non-parametric model which is applicable (and could be adaptive) under different process conditions. This could allow for online use in process control. In addition, one only requires information up to a few moments instead of the entire population distribution, which is possible due to the closed moment equations. Overcoming this closure issue might allow to include influence of particle size on crystal growth. Furthermore, the method as presented in this paper could be extended. For example, in secondary nucleation, it is known that the second moment also affects the total nucleation dynamics, which could be included by extending the covariance kernel k_B for 2-dimensional input accordingly.

REFERENCES

Bari, A.H. and Pandit, A.B. (2018). Sequential crystallization parameter estimation method for determination of nucleation, growth, breakage, and agglomeration kinetics. *Industrial and Engineering Chemistry Research*, 57, 1370–1379.

Deisenroth, M.P., Faisal, A.A., and Ong, C.S. (2020). *Mathematics for Machine Learning*. Cambridge University Press. URL `https://mml-book.com`.

Kostoglou, M. and Karabelas, A.J. (2005). On the self-similar solution of fragmentation equation: Numerical evaluation with implications for the inverse problem. *Journal of Colloid and Interface Science*, 284, 571–581.

McHutchon, A. and Rasmussen, C. (2011). Gaussian process training with input noise. *Advances in Neural Information Processing Systems*, 24, 1341–1349.

Myerson, A.S. (2002). *Handbook of Industrial Crystallization*. Elsevier, 2nd edition.

Porru, M. and Özkan, L. (2017). Monitoring of batch industrial crystallization with growth, nucleation, and agglomeration. Part 1: Modeling with method of characteristics. *Industrial and Engineering Chemistry Research*, 56, 5980–5992.

Ramkrishna, D. (2000). *Population Balances: Theory and Applications to Particulate Systems in Engineering*. Academic Press, 1st edition.

Ramkrishna, D. and Singh, M.R. (2014). Population balance modeling: Current status and future prospects. *Annual Review of Chemical and Biomolecular Engineering*, 5, 123–146.

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian processes for machine learning*. MIT Press. URL `http://www.gaussianprocess.org/gpml/`.

Särkkä, S. (2011). Linear operators and stochastic partial differential equations in Gaussian process regression. In *21st International Conference on Artificial Neural Networks*, 151–158. Springer.

Savvopoulos, S.V., Hussain, M.N., Jordens, J., Waldherr, S., Gerven, T.V., and Kuhn, S. (2019). A mathematical model of the ultrasound-assisted continuous tubular crystallization of aspirin. *Crystal Growth and Design*, 19, 5111–5122.

Appendix A. METHOD OF CHARACTERISTICS

The PBE in (7) describes the NDF as function of time t and particle size x. Once (9) is solved, one obtains the concentration $C(t)$ as function of time which permits the PBE to be solved using the method of characteristics.

In this method, the NDF is solved along a characteristic trajectory $(t(s), x(s))$ as function of a parameter s, where the trajectory starts at some location θ positioned either on the initial condition, IC with $t = t_0$, or on the boundary condition, BC with $x = 0$. The solution along the characteristic trajectory is described by

$$\frac{d\tilde{n}(s,\theta)}{ds} = \frac{\partial n(t,x)}{\partial t}\frac{dt}{ds} + \frac{\partial n(t,x)}{\partial x}\frac{dx}{ds} = -d\tilde{n}(s,\theta), \quad \text{(A.1)}$$

with $\tilde{n}(s,\theta) = n(t,x)$. This equation holds if

$$\frac{dt(s)}{ds} = 1 \text{ and } \frac{dx(s)}{ds} = G(C(t(s))). \quad \text{(A.2)}$$

The solution of the NDF $\tilde{n}(s,\theta)$ in terms of characteristic parameters has a simple analytical solution, from solving the ODE in (A.1),

$$\tilde{n}(s,\theta) = \tilde{n}(0,\theta)\exp(-ds). \quad \text{(A.3)}$$

The problem is now to obtain an inverse expression for s and θ as function of t and x, or in other words, describe the shape of the characteristic trajectories in the $t-x$ plane. For this purpose, one needs to distinguish solutions originating from the IC, and solutions starting from the BC. Figure A.1 offers a visual interpretation of the characteristic trajectories.

First consider a solution starting from the IC, then for $s = 0$, one has $t = t_0$ and $x = \theta \geq 0$. Integration of (A.2) results in

$$\begin{aligned} s &= t - t_0, \\ \theta &= x - \int_{t_0}^{t} G(C(\tau))d\tau, \\ \tilde{n}(0,\theta) &= n_0(\theta). \end{aligned} \quad \text{(A.4)}$$

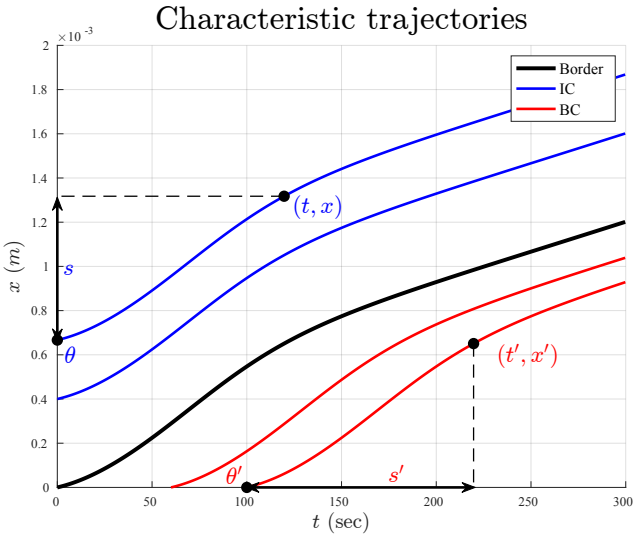


Fig. A.1. Graphical representation of characteristic trajectories starting at IC (blue) and BC (red), alongside with the parametrization in the points (t,x) at IC and (t',x') at BC.

Second, consider solutions starting from the BC, in which case $t = \theta \geq t_0$ and $x = 0$ for $s = 0$. Similarly as to (A.4), one obtains from integration of (A.2)

$$\begin{aligned} s &= \sigma \\ \theta &= t - \sigma, \\ \tilde{n}(0,\theta) &= \dot{n}_0(\theta) = \frac{B(C(\theta))}{G(C(\theta))}, \end{aligned} \quad \text{(A.5)}$$

where σ is the solution of the Volterra integral equation

$$x = \int_{t-\sigma}^{t} G(C(\tau))d\tau. \quad \text{(A.6)}$$

Although not encountered in this paper, in case $G(C)$ switches sign during the integration interval (which would indicate a switch between crystals growing/dissolving), multiple solutions for σ exist. In this case, the largest solution for σ should be chosen for which the characteristic trajectory is confined within the boundary of x for all $s \leq \sigma$. In other words, the trajectory must not cross the boundary condition $x = 0$.

The equations in (A.4) and (A.5) provide an inverse transformation for the characteristic trajectories, which can be substituted into (A.3). What remains, is formulating a condition to determine whether a pair (t,x) lies on a characteristic trajectory starting from the IC or BC. In case that

$$\int_{t_0}^{t} G(C(\tau))d\tau < x, \quad \text{(A.7)}$$

then the IC should be considered, otherwise the BC. An exception is in the case the two sides in (A.7) are strictly equal. This corresponds with the trajectory starting from $(t_0, 0)$, such that both IC and BC apply. In this case, the solution along this trajectory may be ill-defined if the IC and BC differ; the trajectory lies at the edge of a discontinuity.