ENTROPIC CRITERIA FOR COMPUTATIONAL MODELS OF ADVECTION-DIFFUSION EQUATIONS

by

Nhat Thanh Van Tran

A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Master of Science (Applied Mathematics and Statistics).

Golden, Colorado

Date _____

Signed: _____
Nhat Thanh Van Tran

Signed: _____
Dr. Stephen Pankavich
Thesis Advisor

Golden, Colorado

Date _____

Signed: _____
Dr. Greg Fasshauer
Professor and Head
Department of Applied Mathematics and Statistics

ABSTRACT

Traditional probabilistic methods for the estimation of parameters within advection-diffusion equations (ADEs) often overlook the entropic contribution of the discretization, i.e. number of particles, within associated numerical methods. Many times, the gain in accuracy of a highly discretized numerical model is outweighed by its associated computational costs. The research project herein seeks to answer the question of how many particles one should use in a numerical simulation to best approximate and estimate parameters in one-dimensional advective-diffusive transport with constant coefficients. To answer this question, we use the well-known Akaike Information Criteria (AIC) and a recently-developed correction called the Computational Information Criteria (COMIC) to guide the model selection process. Two Lagrangian numerical methods - the random-walk particle tracking (RWPT) and mass-transfer particle tracking (MTPT) methods - are employed to solve the ADE at various levels of discretization. The numerical results demonstrate that the newly developed COMIC provides an optimal number of particles that can describe a more efficient model in terms of parameter estimation and model prediction compared to the model selected by the AIC. These results demonstrate the need for future modelers and scientific researchers to utilize computationally-driven selection criteria in order to best select numerical models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS

Bin size . . . . . Δx

Concentration data . . . . . $\hat{c}_1, \ldots, \hat{c}_k$

Concentration function . . . . . $c(x,t)$

Concentration numerical approximation . . . . . $c_n(x,t)$

Diffusion Coefficient . . . . . D

Final Time . . . . . T

Number of bins . . . . . N

Number of data points . . . . . k

Number of parameters . . . . . p

Number of particles . . . . . n

Temporal step size . . . . . Δt

Velocity Coefficient v

ACKNOWLEDGMENTS

I cannot express enough gratitude to my advisor, Dr. Stephen Pankavich for his continued support and encouragement. I also wish to thank him for his careful review and editing of my thesis. Furthermore, I am grateful to my co-advisor, Dr. David Benson. His knowledge in the field of hydrology has had an immense impact on my work. I offer my sincere appreciation for the learning opportunities and helpful discussions provided by both of my advisors. In addition, I would like to thank my committee members: Dr. Luis Tenorio and Dr. Karin Leiderman, for their insightful comments and questions. Last but certainly not least, I would like to thank my parents, Theo Van Tran and Cuc Thi Pham, for their support and sacrifice over the years to help me pursuing higher education.

For those that shall follow after.

CHAPTER 1
INTRODUCTION

Most realistic particle simulations to solve advection-diffusion problems arising in hydrology use large numbers of particles (often on the order of $10^6 - 10^9$ particles) to fit field data and resolve, with a high degree of accuracy, the fine details of a chemical concentration. However, what such models gain in precision, they typically lack computational efficiency. For practical reasons, this is a crucial issue as useful computational models must find a proper balance between accuracy, parsimony, and efficiency. Within the following chapters, we will examine an information-theoretic criterion that serves to address this issue and answer the following questions:

1. What is the "optimal" discretization for a given numerical method, i.e. number of particles or the grid size, when considering accuracy, complexity, storage, and overfitting concerns?

2. How do the results of simulations using "optimal" models compare to those of more accurate models, in term of parameter estimation and model prediction?

The results herein will demonstrate the need for future researchers in the hydrological sciences to utilize computationally-driven model selection criteria to choose numerical models and physical parameters that more accurately and efficiently describe data collected from advection-diffusion systems. In the next chapter, we will discuss the partial differential equation (PDE) model that is used to describe the hydrological problem. This portion also includes descriptions of two distinct particle tracking methods - random walk (RWPT) and mass transfer (MTPT) - that are used to simulate the PDE and obtain numerical results, as well as, the imbedded statistical model that arises from the need to incorporate data. In Chapter 3, we will expand the methods to more realistic data sets (rather than simulated

data) and discuss the effects of such alterations to the numerical method. Chapter 4 will conclude with the extension of our numerical results to other scenarios in which there is a lack of data or the collected data is sparsely distributed throughout the physical domain (i.e., does not arise from a uniformly-spaced sample). Our ongoing conclusions are summarized at the end of each chapter. As a final point regarding notation, we mention that throughout this work we will use $\log(n)$ to represent the base-10 logarithm instead of $\log_{10}(n)$, while $\ln(n)$ will denote the base-e logarithm.

CHAPTER 2

PROBLEM FORMULATION, MODEL, AND NUMERICAL METHODS

In this chapter, we will describe the physical problem under study, an associated diffusive model of this phenomena, the error incurred both within our theoretical approximation and due to measurement error in collected data, and finally, the underlying PDE-based statistical regression model.

2.1 Physical Problem

The fundamental problem of interest herein comes from hydrology; in particular, we wish to understand the spreading behavior of a contaminant. The problem arises as follows: a source of contamination is introduced into an aquifer or other body at a single spatial point and some initial time. The contamination diffuses and moves throughout the aquifer due to the velocity field of an inherent current. Our main interest is to reconstruct the concentration of the contaminant at any given spatial point within the aquifer and future time. This behavior can be well described using the advection-diffusion equation:

$$\frac{\partial c}{\partial t} = -\nabla \cdot (\mathbf{v}c) + \nabla \cdot (\mathbf{D}\nabla c). \tag{2.1}$$

Here, t is time, $x \in \Omega$ is the spatial variable, $\mathbf{v}$ is a velocity vector, $\mathbf{D}$ is a (symmetric, positive definite) diffusion tensor, $\Omega \subset \mathbb{R}^3$ is the domain of interest, and $c(x,t)$ is the unknown concentration function. Corresponding to this partial differential equation describing the concentration is a system of (Ito) stochastic differential equations that represent the spatial motion of contaminant particles within the aquifer, given by

$$d\mathbf{X}_t = \mathbf{v}\, dt + \sqrt{2}\mathbf{B}\, d\mathbf{W}_t. \tag{2.2}$$

In this equation $\mathbf{X}_t$ represents a vector-valued stochastic process that models particle positions, $\mathbf{W}_t$ is three-dimensional Brownian motion, and the matrix $\mathbf{B}$ is defined via a Cholesky decomposition, namely $\mathbf{B}\mathbf{B}^T = \mathbf{D}$, of the diffusion tensor [1]. In a standard implementation

of an Euler numerical approximation with a specific values of drift and diffusion terms, the Itô equation of particle motion is discretized in time, yielding

$$\Delta X = \mathbf{v}\Delta t + \mathbf{B}\sqrt{2\Delta t}\zeta \tag{2.3}$$

where ζ is a three-dimensional vector of independent standard normal random variables [2].

In the simplest case, this problem is set in one spatial dimension with constant diffusion and velocity terms so that $\mathbf{D} = D$ and $\mathbf{v} = v$, and the PDE model (2.1) can be expressed as

$$\frac{\partial c}{\partial t} = -v\frac{\partial c}{\partial x} + D\frac{\partial^2 c}{\partial x^2}. \tag{2.4}$$

The Euler approximation of (2.2) equation then simplifies to

$$\Delta X = v\Delta t + \sqrt{2D\Delta t}\zeta. \tag{2.5}$$

Because the contaminant is assumed to have originated from a single point, we further take the initial condition to have the special form of a Dirac-delta distribution at an unknown point $x_0 \in \Omega$ so that

$$c(x = x_0, t = 0) = \delta(x_0). \tag{2.6}$$

Given a suitable boundary condition

$$\lim_{x \to \pm\infty} c(x, t) = 0. \tag{2.7}$$

Though it is a classical result, we remark that this problem has a known analytical solution when $\Omega = \mathbb{R}$, namely

$$c(x, t) = \frac{1}{\sqrt{4\pi D t}} \exp\Big[\frac{-(x - vt - x_0)^2}{4Dt}\Big], \tag{2.8}$$

and this will serve as a guide in approximating solutions to the problem posed on a bounded domain in future sections.

2.2 Incorporation of Error from Data and Discretization

Though this PDE model generally provides a realistic estimate of the behavior of the contaminant, the parameters v and D cannot be known *a priori*. Instead, we must infer the values of these parameters from collected data at a specific time $T > 0$. Additionally, as an explicit functional form is not known for solutions of the ADE on a bounded domain, we must instead take a computational approach to representing solutions. Each of these processes will possess intrinsic error and we must devise new methods of quantifying and overcoming the limitations imposed by these inaccuracies.

In order to fit parameter values, we first assume that accurate data is collected at k spatial points within Ω and provided to us. The set of spatial values at which data is collected is represented by the set $\{x_1, ..., x_k\}$, while the corresponding concentrations are denoted by $\{\hat{c}_1, ..., \hat{c}_k\}$. As before, we let $c(x, t)$ represent the solution of the ADE, so that the intrinsic data model can be expressed by

$$\hat{c}_j = c(x_j, T) + \epsilon_j, \tag{2.9}$$

for $j = 1, ..., k$ where $\epsilon_j \sim N(0, \sigma^2)$ are normally-distributed random variables that represent the error between the measured data and the exact solution arising from the model and σ^2 is an unknown variance. These random variables are incorporated within our description of the system in order to account for measurement error that typically arises in the data collection process. Of course, since we don't actually know $c(x, t)$, we must further determine methods to precisely estimate this quantity. Hence, we will approximate the exact solution $c(x, t)$ by a numerical approximation from a computational method, the solution of which depends upon another parameter $n > 0$. For our purposes, we will typically utilize Lagrangian particle methods so that n will be the number of particles in a simulation, but the same can be done for Eulerian methods, in which case $n = 1/\Delta x$ may represent the inverse of the step size. Regardless of the approximation method, we will denote the numerical solution of the ADE by $c_n(x, t)$, and thus the numerical model can be written as

$$c(x,t) = c_n(x,t) + \epsilon(n), \tag{2.10}$$

where $\epsilon(n)$ is a deterministic function representing the error incurred by numerical approximation. Combining equations (2.9) and (2.10), we ultimately arrive at the PDE-embedded statistical model

$$\hat{c}_j = c_n(x_j, T) + \epsilon_j + \epsilon(n) = c_n(x_j, T) + \epsilon_j(n), \tag{2.11}$$

for $j = 1, ..., k$ where $\epsilon_j(n)$ is the combination of error between the data, the exact solution of the PDE model, and the numerical approximation.

Of course, the introduction of statistical measurement error further complicates the situation, as we have instituted an unknown variance $\sigma^2 > 0$ to parametrize the distribution of the errors between observations and model. However, a well-known method for determining an estimator for this quantity exists, and we will discuss this within the following section.

2.3 Maximum Likelihood Estimator (MLE) for σ

To begin, we first direct the reader to Hill and Tiedeman [3] and Brockwell and Davis [4] for references of MLEs used to obtain parameter estimates in models with unknown structure. Under the assumption that the errors between the model and observations are independent, zero-mean Gaussians, the likelihood function is given by

$$L(y;\theta) = \left[(2\pi)^k |\Sigma(\theta)|\right]^{-1/2} \exp\left(-\frac{1}{2} y^T \Sigma(\theta)^{-1} y\right), \tag{2.12}$$

where k is the number of observation points, $\Sigma(\theta)$ is a covariance matrix of errors that depends upon some unknown parameter vector θ, and y is a vector of residuals satisfying $y_j = \hat{c}_j - c(x_j, T)$ for $j = 1, ..., k$. Recall that $\hat{c}_j$ is the measured concentration and $c(x_j, T)$ represents the concentration at the spatial data point x_j and time T given by the PDE solution. Therefore, the associated log-likelihood function is

$$\ln(L) = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}|\Sigma| - \frac{1}{2} y^T \Sigma^{-1} y. \tag{2.13}$$

As in the situation we've stated in the previous section, the observation errors are often assumed to be of this form, and additionally, Σ is diagonal. Furthermore, the variance of each observation is often unknown or estimated during the model regression (e.g., see Chakraborty et al. [5] for derived concentration errors in particle methods), so it is assumed that Σ depends only upon a single variance parameter, denoted by σ^2, and thus satisfies $\Sigma = \sigma^2 \mathbb{I}$. The last term in equation (2.13) is more conveniently given in terms of the sum of squared errors

$$\text{SSE} = y \cdot y = |y|^2$$

so that

$$\ln(L) = -\frac{k}{2}\ln(2\pi) - \frac{k}{2}\ln\sigma^2 - \frac{k}{2\sigma^2}\frac{\text{SSE}}{k}. \tag{2.14}$$

Because this function should be maximized, we attempt to compute the roots of the derivative of $\ln(L)$ with respect to σ^2 in order to identify the value(s) of σ^2 at which maximums occur. Doing so provides an estimator of the observation variance, namely $\hat{\sigma}^2 = \text{SSE}/k$, so that the corresponding maximum value of the log-likelihood is

$$\ln(L) = -\frac{k}{2}\left(1 + \ln(2\pi) + \ln\left(\frac{\text{SSE}}{k}\right)\right). \tag{2.15}$$

Since the number of observations is usually fixed, the $\frac{k}{2}$ term is constant and can thus be canceled from all terms. Similarly, the remaining constants do not change from one model evaluation to another, and can also be omitted. Hence, the MLE is $\hat{\sigma}^2 = \text{SSE}/k$ under these assumptions, and the quantity $-\ln(\text{SSE}/k)$ provides a relative estimate for the value of the log-likelihood function evaluated at this MLE.

Summarizing the derivation of the maximum likelihood estimator for the variance, we can now precisely express the statistical-PDE model of the previous section as

$$\hat{c}_j = c_n(x_j, T) + \epsilon_j(n), \tag{2.16}$$

for $j = 1, ..., k$ where

$$\epsilon_j(n) = \epsilon_j + \epsilon(n)$$

is the error in our approximation and $\epsilon_j \sim N(0, \hat{\sigma}^2)$ is determined by the estimator $\hat{\sigma}^2$, which satisfies

$$\hat{\sigma}^2 = \frac{1}{k}\sum_{j=1}^{k} |\hat{c}_j - c_n(x_j, T)|^2 = \frac{\text{SSE}}{k}, \tag{2.17}$$

as $c_n(x_j, T)$ is our best approximation of the unknown value $c(x_j, T)$.

2.4 Computational Information Criterion (COMIC)

With the numerical model established, we first note that we have actually defined an entire collection of such models, each of which depends upon the choice of the computational parameter n. Hence, we can now turn our attention to identifying a criterion for selecting the "best" model among these. One guide is Akaike's information criterion (or AIC), which presents a criterion based on entropy maximization that allows one to select from a set of models depending upon an unknown number of undetermined parameters. Though our particular model does possess two distinct parameters (D and v), we do not technically allow for the number of parameters to vary here. Instead, the issue is selecting amongst a different, computational, number of parameters. Therefore, the ultimate goal of this section is to derive an extension of the AIC that establishes an objective function to be optimized in order to select a model and a minimal number of parameters that best fits a given set of data [6]. This discussion may require some background knowledge of the Kullback-Leibler divergence (KLD) and the basic formulation of the AIC, which can be found in recent work by Benson et al [7].

We will start with a summary of the entropy calculation. For a probability mass function $p(x)$ taking non-zero values at points $x_1, \ldots, x_k$, the discrete entropy is defined by

$$H_D(X) = -\sum_{i=1}^{N} p(x_i)\ln(p(x_i)). \tag{2.18}$$

It follows that the entropy for a probability density function $f(x)$ is analogous to the discrete case and listed as

$$H_I(X) = -\int_{f(x)>0} f(x)\ln(f(x))dx. \tag{2.19}$$

However, equation (2.19) is not well defined because $f(x)$, i.e. the argument of ln(), is not dimensionless. In order to calculate the entropy of a PDF, we will impose a sampling interval $\Delta V > 0$. For a small ΔV, an entropy H_C can be defined that is consistent with H_D using

$$\mathbb{P}(x - \Delta V/2 < X < x + \Delta V/2) \approx f(x)\Delta V, \tag{2.20}$$

so that the equation (2.19) becomes

$$\begin{aligned} H_C(X) &= -\int_{f(x)>0} f(x)\ln(f(x)\Delta V)dx \\ &= -\ln(\Delta V) + H_I. \end{aligned} \tag{2.21}$$

Therefore, H_C and H_D are comparable values, and we have

$$H_D \approx H_C = -\ln \Delta V + H_I. \tag{2.22}$$

Next, we would like to use the KLD to measure the relative entropy of the error distribution between the observed data and numerical approximation of the PDE. The crucial point here is that we must adjust the criterion to account for the fact that the discrete approximation depends upon the resolution of the numerical method, which is a function of the single parameter n. In the current context, this parameter represents the number of particles n in a stochastic particle method, but it can just as easily identify the spatial grid size, say $\Delta x = |\Omega|/n$, in a finite difference method. Ultimately, we wish to establish a criterion to select the best of these approximate models depending upon the value of n.

Let $(x_j, \hat{c}_j)$ for $j = 1, ..., k$ represent the given pairs of concentration data and $c(x, t)$ denote the true solution to the PDE model, we can describe the statistical model incorporating measurement error by

$$\hat{c}_j = c(x_j, T) + \epsilon_j$$

for $j = 1, ..., k$, where T is a known measurement time and $\epsilon_j \sim h(y_j|\theta)$ is a random variable with distribution h that encodes each of the associated random errors.

It is difficult to obtain an analytic solution $c(x, T)$ for realistic problems, which means that we must approximate the PDE solution at $t = T$ with a number of suitable numerical models, the solutions of which we denote by $c_n(x, T)$. Because there will be discrete approximation error between the numerical approximation $c_n(x, T)$ and the analytic solution $c(x, T)$, we must incorporate the implications of the discrete approximation within the model selection criterion. Fortunately, the KLD can account for the latter quantity, and we can establish a computational information criteria by adjusting the AIC by the difference in entropy between the numerical approximation and the analytic solution. If we let $H_{rel}(f_1, f_2)$ represent the relative entropy between the distributions f_1 and f_2, then the new information criterion can be expressed as

$$\text{COMIC}(\hat{\theta}, n) = H_{rel}(c(\hat{\theta}), \hat{c}) + H_{rel}(c_n(\hat{\theta}), c(\hat{\theta})).$$

The first term is given by the AIC, which is

$$\text{AIC} = -2\sum_{j=1}^{n} \ln h(y_j|\hat{\theta}) + 2p \tag{2.23}$$

while the second term is the difference between the discrete entropy of the numerical approximation, and the inconsistent entropy of the analytic solution, or $H_D - H_I$. From equation (2.22), we have

$$H_D - H_I \approx -\ln(\Delta V).$$

Using this, we can define an adjusted criterion to the AIC, which we name the COMIC or the COMputational Information Criteria, given by

$$\text{COMIC}(\hat{\theta}; \Delta V) = -2\sum_{j=1}^{n} \ln h(y_j|\hat{\theta}) + 2p - \ln(\Delta V) = \text{AIC} - \ln(\Delta V). \tag{2.24}$$

This computational criteria can be interpreted as a limitation on the information content needed to represent the approximate solution $c_n(x, t)$. Given suitably rapid convergence properties of a numerical method, the value of $c(x, T)$ can be computed to an arbitrarily

large degree of precision by merely choosing n sufficiently large. However, in doing so, one must use an increasingly prohibitive amount of information in order to gain greater levels of accuracy. Thus, there is a diminishing return between the desired level of precision and the required information content. In this way, the COMIC is a criterion for penalizing such considerations to select a parsimonious and computationally efficient (i.e., low information content) model.

Next, we will apply this criterion to the problem established within the previous section, in which the errors between model and observations are Gaussian with variance σ^2, and assume that no other parameters (e.g., D or v) require estimation. In this scenario, the log-likelihood function evaluated at the maximum-likelihood estimate is proportional to the log of the average sum of squared errors (SSE) given by eq. (2.15). Upon taking the sampling volume to be the average particle spacing within the interval, namely $\Delta V = \frac{|\Omega|}{n}$, and removing constants, the form of the COMIC becomes

$$\mathrm{COMIC}(n) = 2\ln\left(\frac{\mathrm{SSE}}{k}\right) + \ln(n), \tag{2.25}$$

where

$$\mathrm{SSE} = \sum_{j=1}^{k} |\hat{c}_j - c_n(x_j, T)|^2. \tag{2.26}$$

For reference, the AIC without a parameter correction term, as described above, is merely

$$\mathrm{AIC} = 2\ln\left(\frac{\mathrm{SSE}}{k}\right). \tag{2.27}$$

With this representation (2.25) of the COMIC, we can easily see that numerical models with equivalent SSE - meaning that their measure of distributional entropy is the same - will possess a computational entropy difference of $\ln(n)$, implying that the model fitness must be adjusted by this contrast in their underlying information content.

2.5 Particle Methods

For a standard stochastic particle method to construct numerical solutions of equation (2.1), a large number of particles are transported according to the computational implementation of the stochastic differential equation (2.3) for particle motion. Then, the histogram of particles at some fixed time $T > 0$, along with a chosen binning method, is used to estimate the solution $c(x, t)$. In addition to this classical, random-walk method, we will also investigate and utilize a newer mass-transfer method, which can be implemented without randomness.

2.5.1 Random Walk Particle Tracking

In the random walk particle tracking (RWPT) method, the equation (2.2) can be implemented directly. Given a particle at position $X(t)$ at time t, the position of the particle is advanced to the next time step using the first-order Euler approximation

$$X(t + \Delta t) = X(t) + v\Delta t + \sqrt{2D\Delta t}\zeta \tag{2.28}$$

as in equation (2.3), where ζ is a standard normally-distributed pseudo-random number and $\Delta t > 0$ is a chosen time step. Of course, $X(0)$ is determined by an initial particle spacing, and using (2.28) this processes is repeated until the final time step $T = N\Delta t$ for some number of steps N. At the final time step, the particle positions are all recorded, and a selected binning method is used to place particles into intervals with a predetermined length, thereby reconstructing the concentration.

More precisely, we can describe the iterative steps of the numerical method as follows:

1. Define the n particle positions at time $t = 0$ to be $X(0) = x_0$, each possessing a mass of $1/n$. In this way, the total mass is preserved (to be 1) by the method and the contaminant particles each begin at the same position x_0.

2. At each time step, every particle is transported according to the equation (2.28).

3. At the final time step, the position and mass of the particles is used to compute the estimated concentration with a binning method (to be described in greater detail below).

2.5.2 Mass Transfer

In addition to the RWPT, we utilize a mass transfer particle tracking (MTPT) method to approximate solutions of the ADE. In this method, each particle is fixed at a uniformly-spaced position, and instead of moving according to the Ito equation (2.2), the mass amongst particles is transferred according to their probability of collocation. Thus, instead of binning the particles and performing ensemble averages to compute the concentration, we merely change the values of the concentration according to the aforementioned probability. In this way, the MTPT method is similar to a standard finite-difference method, but the rule used to transfer mass is not based upon a Taylor expansion; rather a discretization of the convolution of the initial condition with an approximate fundamental solution is utilized to determine how mass is transferred amongst particles. Of course, this can also be done using a first-principles physical derivation as in [8].

Due to equation (2.3), we know that at each time step, the probability of a given particle existing at a certain location is normally distributed with mean given by its current particle position and a standard deviation of $\sqrt{2D\Delta t}$. In this way, we can intuitively express each particle position according to the normal distribution instead of a Dirac-delta distribution. This means that for a pair of particles $X_i(t)$ and $X_j(t)$ their shared probability density of collocation is given by

$$P_{ij} = \sqrt{\frac{\Delta s}{8\pi D_{ij}\Delta t}} \exp\Big(- \frac{r_{ij}^2}{8D_{ij}\Delta t}\Big), \tag{2.29}$$

where Δs is a particles support volume used to normalize the discrete probability density, D_{ij} is the average value of the diffusion tensor at the ith and jth particles, and r_{ij} is the distance between the i and j particles. For the problem of interest here, D_{ij} is constant ($D_{ij} = D$), which simplifies equation (2.29). Additionally, since we wish to enforce the condition that

for each $i = 1, ..., n$ the vector p_i defined entrywise for each $j = 1, ..., n$ by $(p_i)_j = P_{ij}$ is a probability mass function, we further impose the normalization

$$\sum_{j=1}^{n} P_{ij} = 1 \tag{2.30}$$

for all $i = 1, 2, \ldots, n$, and this fixes the value of Δs so that it does not affect our calculations. When the n probability mass functions are assembled into the matrix P, this constraint is merely row normalization, but it does not guarantee that the columns also sum to one. In general, we would like to preserve the symmetry of this matrix and to do this while also enforcing an approximate normalization, we use the average row and column normalization to construct the symmetric and probability mass preserving matrix P. With the probability matrix constructed, we allow the masses to exchange between pairs of particles, and as a result, the first order approximation of the mass of the i^{th} particle at the next time step is given by

$$m_i(t + \Delta t) = m_i(t) + \sum_{j=1}^{n} \frac{1}{2}(m_j(t) - m_i(t))P_{ij}. \tag{2.31}$$

In summary, at each time step the mass transfer method allow particles to exchange masses between nearby particles. As can be immediately deduced from (2.31), mass will move from particles with greater mass to those with lesser mass with the additional constraint that their mass difference is split evenly [8], [9]. Though we have provided a more physical description of the interaction, (2.31) can also be interpreted as a discretization of the convolution of the Gaussian densities representing the position of the i^{th} and j^{th} particles.

To implement this numerical method, we perform the following steps:

1. Place the particles within the domain at equally spaced gridpoints, as in a finite difference method.

2. Because our domain is centered at $vT + x_0$, the initial condition may be altered to match the spacing. Hence, depending on the total number of particles, their spacing may differ. For instance, if n is odd, there will be always a particle at $vT + x_0$.

Therefore, we set this particle mass to be 1 to match the given initial condition. If instead, n is even then we approximate the initial condition $c(x_0, 0) = \delta(x_0)$ using the two nearest particles to the center $vT + x_0$ - the one to the left and one to the right of this point are each given a mass proportionate to their respective distance to the center. More specifically, the first order approximation will be used to set the initial mass of particles, and this is given by

$$m_i = 1 - \frac{vT + x_0 - x_i}{r}, \tag{2.32}$$

where r is the distance between these two particles, and i is either the left or the right particle.

3. For each particle, identify all nearby particles and their associated distance using a range search criteria. Given a specific radius around a particular particle, this algorithm will determine all particles within the radius and returns the particle position index and distance between the particles.

4. Construct the collision probability matrix P using the information from the range search step above.

5. At each time step, compute the mass of every particle using the equation (2.31).

6. At the final time step, the position and masses of particles are used to represent the concentration via a binning method discussed within the next section. Because the particles in the MTPT method are stationery, the binning essentially reduces to dividing by the average particle spacing; namely length of the domain divided by the number of particles.

Notice that this method integrates the constant velocity v be imposing a translation of the domain. Additionally, as no stochastic motion of particles is implemented, the approximate solution produced from this method is completely deterministic. Because of this, the list

of nearby particles and the distances between them do not change in time, and the range search step need only be implemented one time regardless of the number of timesteps.

2.5.3 Binning Method

The dimension of concentration is $[m/L]$, where m is mass and L is length. In the field of chemistry, the concentration of a substance is often measured by taking an amount of solvent and then drying out the water to obtain a solute. The mass of the resulting solute is then divided by that of the solvent to obtain the concentration. We utilize a similar process to perform appropriate binning of particles within our methods. In particular, we compute the total mass of particles within a certain bin and then divide by the length of the bin to obtain the concentration. As our problem lies within a single spatial dimension, consider $\Omega = [a, b]$. At a given time t, we denote the vector of particle positions by $x(t)$, their corresponding mass vector by $m(t)$, and the vector of equally space binning grid points $b = [b_1, ..., b_{N+1}]$. Thus, the ith bin is exactly $[b_i, b_{i+1}]$, for $i = 1, 2, \ldots, N$, where $b_1 = a$ and $b_{N+1} = b$. Then, the normalized concentration using n particles can be represented as

$$c_n(x,t) = \frac{1}{m_{tot}} \sum_{i=1}^{n} \int_{\Omega} m_i(t)\delta(z - x_i(t))\phi(x,z)\, dz \tag{2.33}$$

$$= \frac{1}{m_{tot}} \sum_{i=1}^{n} m_i(t)\phi(x, x_i(t)), \tag{2.34}$$

where $c_n(x,t)$ is a reconstructed concentration function from n number of particles, m_{tot} is the total mass, $m_i(t)$ is the mass of i^{th} particle, $\delta(x - x_i(t))$ is a Dirac-delta function, and

$$\phi(x, x_i(t)) = \begin{cases} 1/\Delta x, & \text{if } x \in [b_k, b_{k+1}] \\ 0, & \text{else} \end{cases} \tag{2.35}$$

where $k = \left\lceil \dfrac{x_i(t) - b_1}{\Delta x} \right\rceil$ is the binning grid point to the left of the particle position, $\lceil x \rceil$ is the ceiling function of x (so that $\lceil x \rceil = 1 + \lfloor x \rfloor$ where $\lfloor x \rfloor$ is the integer part of x), and Δx is the bin size. Notice that if $x_i(t) = b_1$, then this definition chooses the value of k to be 1 in the numerical simulation, which avoids any out of bounds errors.

Unfortunately, we can not use this exact formulation to find the concentration within a numerical simulation because if the value of $x_i(t)$ is identical to b_k for some $k = 1, ..., N+1$, then the estimated concentration is biased towards a certain bin. For example, if we estimate the concentration at every single binning grid point, the approximate solution will be shifted to the left by the value of the bin size. This is especially problematic when estimating the velocity using the approximate solution; in particular, the estimated velocity will always be greater than the actual value. To avoid this problem, we introduce two types of specialized numerical binning. What we deem the left binning method takes

$$\phi(x, x_i(t)) = \begin{cases} 1/\Delta x, & \text{if } x \in [b_k, b_{k+1}) \\ 0, & \text{else,} \end{cases} \tag{2.36}$$

while the right binning method uses

$$\phi(x, x_i(t)) = \begin{cases} 1/\Delta x, & \text{if } x \in (b_k, b_{k+1}] \\ 0, & \text{else.} \end{cases} \tag{2.37}$$

We can see that each of these binning rules has a problem handling the end point. With the left binning, this occurs when the evaluated particle position is equal to the final grid point, and correspondingly the first binning grid point for the right binning method. To resolve this problem, we introduce an extra bin to capture the end point for each method. Then, we calculate the average of the approximate solutions from both binning methods to obtain our final approximate solution.

2.6 Basic Example with No Parameter Fitting

In this section, we perform a series of introductory simulations for both the RWPT and MTPT methods in order to determine the optimal particle numbers predicted by the AIC and COMIC, respectively. All simulations in this section use $D = 1$, $v = 0$, $x_0 = 0$ and run until the final time $T = 1$. The 30 simulated data points arise the exact solution with equally spaced points on the interval $[-5, 5]$, which is approximately $\pm 3.5\sqrt{2DT}$, covering 3.5 standard deviations of the exact solution.

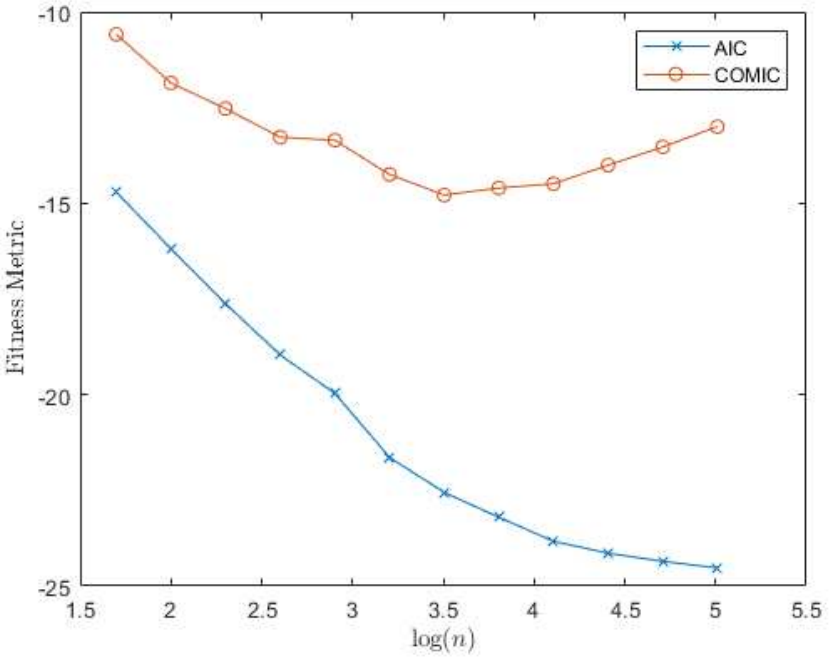

Figure 2.1: Fitness metrics (AIC and COMIC) for the RWPT method with $D = 1$

2.6.1 RWPT Method

For the random walk simulations, we choose the bin size of 0.1 because with this particular bin size, the sum of squared errors is the smallest for this problem regardless of the number of particles. We implemented the simulation with $n = 50 \cdot 2^q$ particles for $q = 0, ..., 11$ powers, calculated the fitness metric, and then repeated this process for a total of 20 times and computed ensemble averages in order to reduce the effect of randomness within each run. From Figure 2.1, we see that the error decreases as the number of particles increases. However, there is a diminishing return on the benefit of using more particles. The COMIC shows that there is an optimal number of particles that balances between the goodness of fit and the computational efficiency of the numerical method, and this occurs at about 3200 to 6400 particles (i.e., $\log(n) \in [3, 4]$). Each run produces different results because of the random nature of the particle motion. However, performing multiple repeated trials and determining the particle number at which the minimum occurs, generally results in the

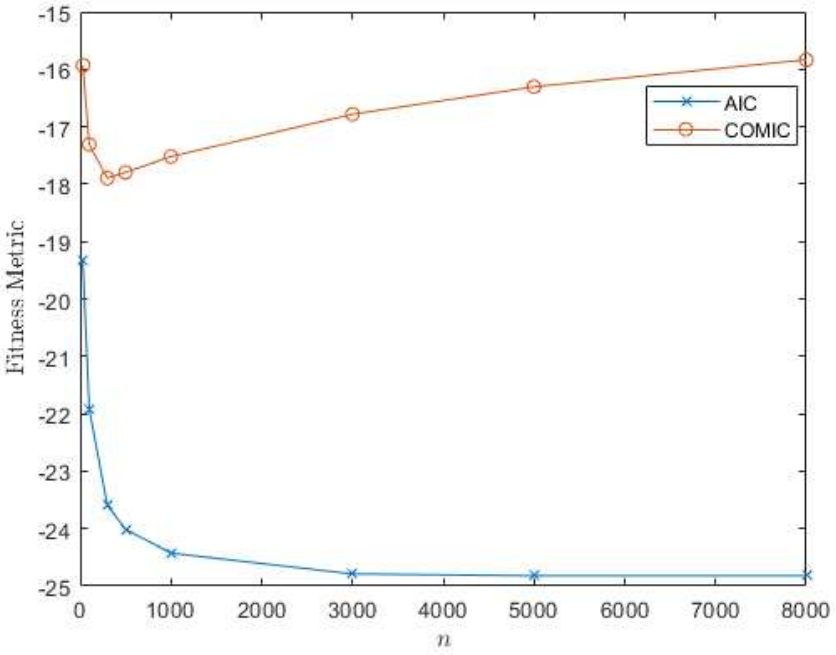

Figure 2.2: Fitness metrics (AIC and COMIC) for the MTPT method with $D = 1$

number of optimal particles lying between $n = 3200$ and $n = 6400$.

2.6.2 MTPT Method

Similar to previous section, we repeat this exercise for the mass transfer method. However, we require a smaller number of particles compared to the random walk method because of its deterministic nature. In this case, we perform the simulation with

$$n = 30, 100, 300, 500, 1000, 3000, 5000, 8000$$

particles. Using the matrix implementation of the method, the approximate solution is deterministic, so simulations need not be repeated, and the bin size is merely the spacing between each particles. From Figure 2.2, we find that the error plateaus very quickly as the number of particles is increased, and the optimal number of particles for this simulation is found to be 300.

2.7 Summary

In this chapter, we discussed the physical problem arising from the hydrological sciences and its corresponding model, which utilizes the scalar advection-diffusion equation. For certain choices of domains, parameters, and initial conditions, this formulation yields an exact solution. Next, we introduce two different particle methods that are used to approximate the solution. The random walk and mass transfer methods demonstrate that they can precisely approximate an exact solution of the ADE, and as the number of particles increases, the numerical solution converges to the exact solution, i.e. the AIC decreases as the number of particles increases. We also examine the COMIC, a new computational information criterion to identify the optimal number of particles that should be used to simulate each method when balancing goodness of fit with computational efficiency. We will use these example simulations as a guideline for those which occur within subsequent chapters.

CHAPTER 3
COMPUTATIONAL RESULTS

In this chapter, we will utilize the theoretical, analytical, and algorithmic development of the previous chapter in order to obtain some computational results concerning the estimation of parameters and the utility of the model selection criterion discussed earlier.

3.1 Parameter Fitting

Using the models outlined in Chapter 2, we can now show demonstrate the efficacy of the proposed fitness metric in the case that the underlying parameter values are known. Of course, in more realistic scenarios, this will not be feasible as the true parameter values cannot be obtained, but instead must be estimated only from the given data. Still, we will verify that the fitness metric ensures an accurate and efficient model given this parameter estimate. We begin by estimating on the diffusion coefficient D. For random walk simulations, the process of estimating D involves the following steps:

1. Formulate the problem with $D = 1$, $T = 1$, $v = 0$, $x_0 = 0$, and an initial guess for the numerical optimization solver of $D_{\text{init}} = 0.5$.

2. Use the AIC and COMIC information criteria to determine the optimal number of particles within a simulation, and then estimate the diffusion coefficient D_{est} with both of these particle numbers. Typically, the COMIC optimal number of particles (denoted n_{COMIC}) will be approximately $n_{\text{COMIC}} = 5000$, while the AIC-predicted number is arbitrarily large. In the latter case, we utilize the maximum number of particles that we can feasibly simulate, namely $n_\infty = 102400$.

3. Repeat this process for 30 trials and create a box plot of all estimated diffusion coefficients.

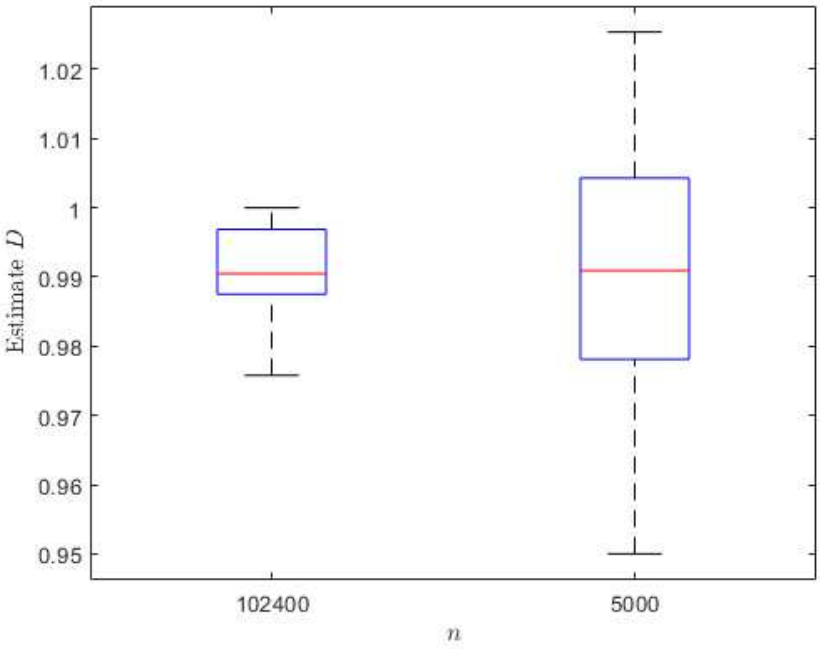

Figure 3.1: Box plot of the random walk diffusion estimate D_{est}.

For the mass transfer simulations, we will use the optimal number of particles as indicated by the COMIC, i.e. $n_{\text{COMIC}} = 300$, and the maximum number of particles of $n_\infty = 8000$. Since this method is deterministic, we need only run the simulation once and record the estimated diffusion coefficient.

The results from the random walk simulations are shown in Figure 3.1. We see that with $n_{\text{COMIC}} = 5000$ particles, the estimated diffusion coefficients ranges from about 0.96 to 1.02, which is within 4% absolute (and relative) error. Comparing this error estimate with that arising from the maximum number of particles $n_\infty = 102400$, which is about 20 times as many particles as n_{COMIC}, the absolute error is about 2.5%. This demonstrates that an enormous number of particles are needed to reduce the error by a fairly marginal about, namely 1.5%. Thus, it is much more efficient to use the COMIC-predicted optimal number of particles to perform the simulation. Similarly, the mass transfer simulation with $n_{\text{COMIC}} = 300$ particles yields an estimated diffusion coefficient of 1.0324, while the maximum number of particles ($n_\infty = 8000$) yields 1.0269. Note that the estimated diffusion coefficients are all greater

than 1, for this particular method because we restrict the particles within a certain domain, meaning that the total mass in that domain does not change, while the exact solution will contain less mass in that same domain as time increases. Of course, the results are quite similar to each other with about a 0.5% difference. This means that utilizing nearly 27 times as many particles results in no significant difference between the estimates. In addition to the gain in computational efficiency, the simulation time required for 8000 particles is notably longer than 300 particles. In particular, a run with 300 particles requires about 0.7 seconds to complete, while a simulation with 8000 particles requires about 7 minutes. For the mass transfer method, it is always better to use the COMIC optimal number of particles, as no significant difference occurs in parameter estimation and the duration of the simulation is much shorter.

Next, we will estimate both parameters, D and v, using the same methodology. In particular, we will set the true velocity parameter to unity, i.e. $v = 1$, and take the initial guess to be $v_{\text{init}} = 0.5$. We do not need to estimate x_0, because the determination of v and x_0 is coupled, meaning that if we know v then x_0 can be determined and vice-versa. This is easily demonstrated by the fundamental solution (2.8).

Figure 3.2 shows similar results as before for the two-parameter study. Within the random walk particle tracking method, the estimated diffusion coefficients are similar to before, and the estimated velocity from $n_{\text{COMIC}} = 5000$ particles displays about 3% error, while the $n_\infty = 102400$ simulations produce an estimate with about 1% error. With regard to the mass transfer method, for $n_{\text{COMIC}} = 300$ particles the estimated diffusion coefficient is 1.0325, just as before, and the estimated velocity is exactly 1 due to the translation of the domain. This means that if one wants to estimate the velocity, the COMIC-produced optimal number of particles will achieve the same result as any greater number of particles. For reference, the estimate of the diffusion coefficient for 8000 particles is 1.0267, while the velocity estimate is 1. Of course, one may suggest that, we do not need to use even $n_{\text{COMIC}} = 300$ particles to estimate the correct velocity. However, a simulation that uses only $n = 150$ particles yields

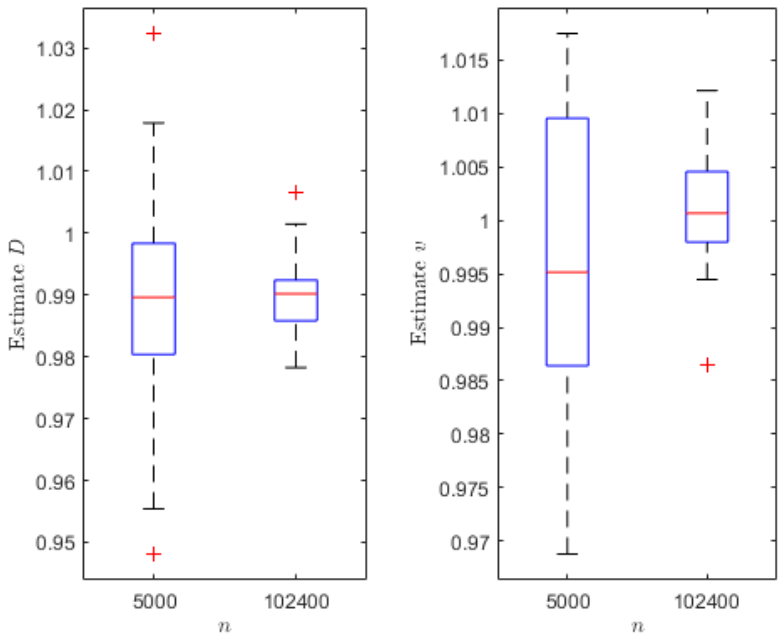

Figure 3.2: Plot of the random walk D and v estimate.

an estimated velocity of 0.9843 and estimated diffusion coefficient of 1.0408. Hence, the $n_{\text{COMIC}} = 300$ simulation provides needed accuracy without an unnecessarily large number of particles.

3.2 Model Selection - COMIC vs AIC

Next, we will compare the performance of the COMIC-generated model (i.e., using n_{COMIC} particles) with the maximum number of n_∞ particles predicted by the AIC. To do so, we will plot the exact solution and the numerical approximation stemming from different particle numbers and quantify the difference amongst these distributions. In Figure 3.3, the results of the random walk particle tracking simulations are displayed. We see that most of the differences between the exact solution, the COMIC numerical model, and the AIC numerical model occur near $x = 0$ where the concentration is expected to be greatest. Of course, the AIC numerical solution better approximates the exact solution, as the goodness of fit portion of this criterion produces a lesser value than that of the corresponding COMIC

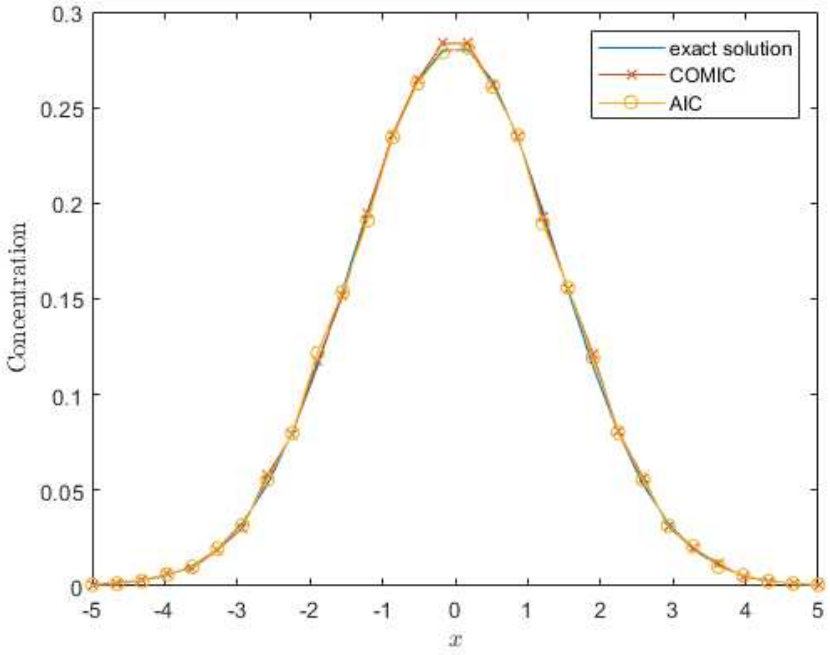

Figure 3.3: Plot of the exact solution and the random walk numerical approximations with different number of particles.

goodness of fit. However, the differences between the COMIC and AIC numerical solutions are quite small. For the mass transfer method shown in Figure 3.4, again most of the errors occur near $x = 0$, and the AIC numerical solution remains closer to the exact solution. However, it is difficult to see the differences between the COMIC and AIC approximations. From this perspective, it is difficult to justify using many more particles (e.g., n_∞) to obtain an approximation that is similar to the COMIC number of particles n_{COMIC}.

3.3 Simulated Data

Naturally, when gathering data in the field, there are many ways a collector can introduce error into the process. Error may arise from limitations of equipment, the specific process in which the data is collected, or via human error. Therefore, even when the underlying distribution of the system comes from the true solution of a mathematical model, in this case the advection-diffusion equation, the data may not stem precisely from the exact solution

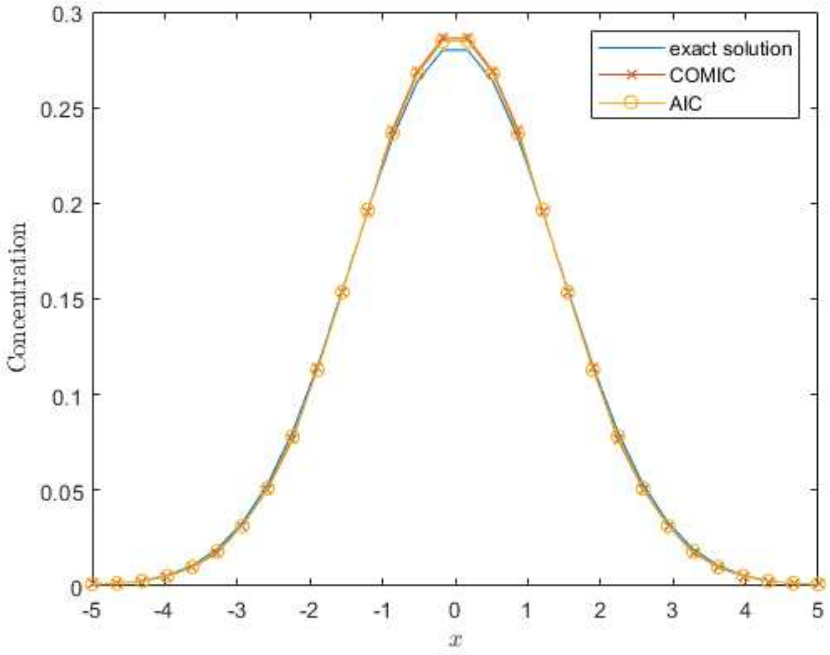

Figure 3.4: Plot of the exact solution and the mass transfer numerical approximations with different number of particles.

to this PDE, i.e. a Gaussian distribution. To increase the versatility and usefulness of the method, the data used to perform these simulations should contain some errors so as to better represent real-world data. In order to accomplish this, we will make some assumptions about the distribution of the statistical error in the model. In particular, we will examine two different error distributions - normal and uniform. To generate the data, we first select a number of data points from the exact solution of the PDE at a fixed time $T > 0$. Then, at each data point, we add a random value from the chosen error distribution. To ensure that the data retains physical meaning, e.g. non-negativity, the standard deviation of these distributions must be carefully chosen. For the normal distribution, we will choose the standard deviation to be 1/3 of the minimal value of the solution on the bounded domain. This does not guarantee that the data will always remain non-negative, but it will do so with exceptionally high probability. We can further resolve this problem by removing unphysical data from a simulation should it occur. To ensure a useful comparison, we will choose

the standard deviation of the uniform distribution to be the same as that of the normal distribution. This implies that the error within the uniform distribution is lower bounded by $-1/\sqrt{3}$ of the minimal value and upper bounded by $1/\sqrt{3}$ of the minimal value. With this particular choice, the data from the assumption of uniformly distributed errors is guaranteed to be non-negative. Also, in order to preserve a large variance of the error distributions, we will select the data from a smaller range of the domain compared to previous simulations, though, the numerical method will continue to use the same domain as before. This means the number of data points will also need to be reduced in order to maintain the uniform spacing of the data.

Numerical trials of the RWPT and MTPT methods with simulated data are performed as follows:

1. Set $D = 1$, $v = 1$, $T = 1$, $x_0 = 0$.

2. With this choice of parameters, the exact solution suggests that the domain of interest will be shifted to the right by one, $[-4, 6]$, in comparison to the original domain $[-5, 5]$.

3. Select 18 equally space data points from the domain $[-2,\ 4]$.

4. Find the minimum value of the collection of data points, namely $\delta = 0.0297$.

5. Construct the error distribution; normally-distributed errors arise from $\epsilon_n \sim N\Big(0, \dfrac{\delta^2}{9}\Big)$, while uniform errors are $\epsilon_n \sim U\Big(-\dfrac{\delta}{\sqrt{3}}, \dfrac{\delta}{\sqrt{3}}\Big)$.

6. For each trial, construct the noisy data using the exact solution and error distribution.

7. Use this noisy data to estimate the diffusion and velocity coefficients, with the initial guess of 0.5 for both, by using the optimal number of particles depending on the numerical method.

8. Repeat this process for 30 trials to account for randomness and average the parameter estimates over these trials.

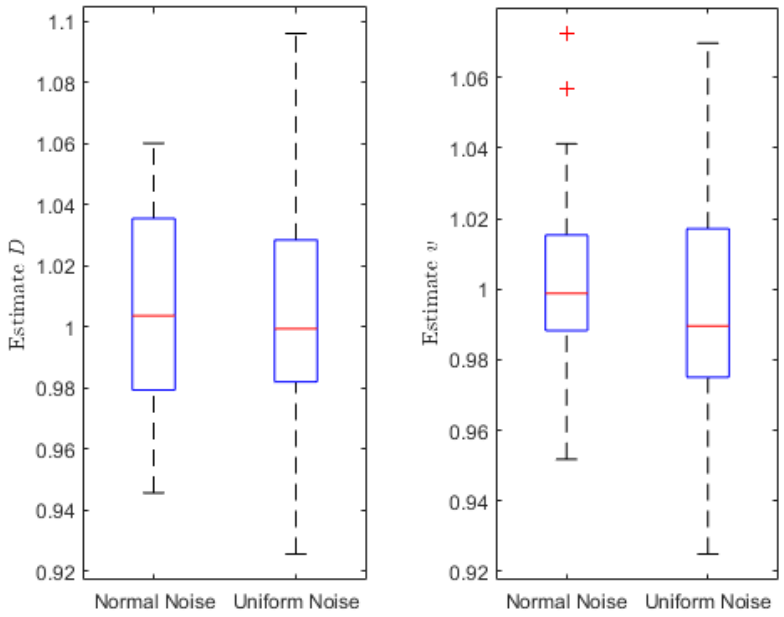

Figure 3.5: Plot of the random walk D and v estimate using noisy data.

With this formulation, we find that the maximum value of the data is 0.2799, while the minimum is approximately 0.0297, which is slightly greater than 10% of the maximum. The random walk results are shown in Figure 3.5. With the normal noise, both coefficients display about 6% error, while with the uniform noise, the error is about $8-10\%$. The mass transfer results are similar, as seen in Figure 3.6.

3.4 ℓ^2 Analysis Between the Exact Solution and Numerical Approximation

Since we know the exact underlying solution of the PDE, we can use the ℓ^2 norm as a metric to verify the COMIC optimal number of particles. Recall that $c_n(x_i,t)$ represents the approximate solution from the numerical method and $c(x_i,t)$ denotes the exact solution, each evaluated at a grid point x_i. Then, the ℓ^2 norm of the difference between the exact versus approximate solution at the specified final time T is

$$\|c-c_n\|_{\ell^2}=\sqrt{\sum_{i=1}^{N}|c(x_i,T)-c_n(x_i,T)|^2}. \tag{3.1}$$

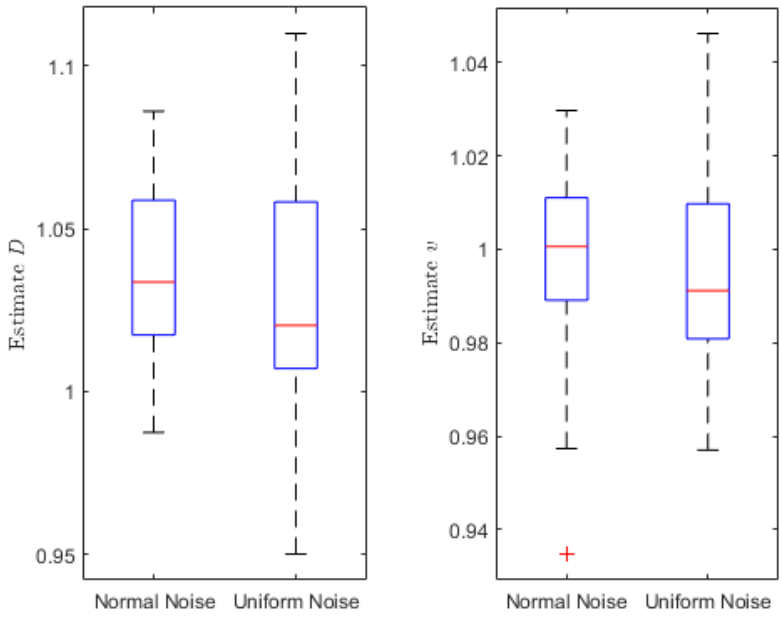

Figure 3.6: Plot of the mass transfer D and v estimate using noisy data.

Notice here that chosen gridpoints $\{x_1, \ldots, x_N\}$ need not be data points from previous simulations. Instead, we may take a much finer discretization of the domain to determine whether, even with noisy data, the underlying distribution, rather than merely its value at specific data points, can be reconstructed. To compute this norm, we perform simulations with $D = 1$, $v = 0$, $T = 1$ and $x_0 = 0$. For the random walk method, we compute the exact solution at each binning point with a bin size of 0.1; hence, there will be 101 such points. For the mass transfer method, we cannot obtain an exact solution at each binning point because the number of bins depends on the number of particles used within a simulation, and this does not allow us compare the same data among different models. We would expect the error to increase as the number of particles increases if this is the case. Beside comparing the error among the number of particles in a given numerical method, we would also like to compare the error between two numerical methods, so we choose the same 101 data points from the random walk method for the mass transfer method. The results of the mass transfer method are shown in Figure 3.8. There is a slight difference

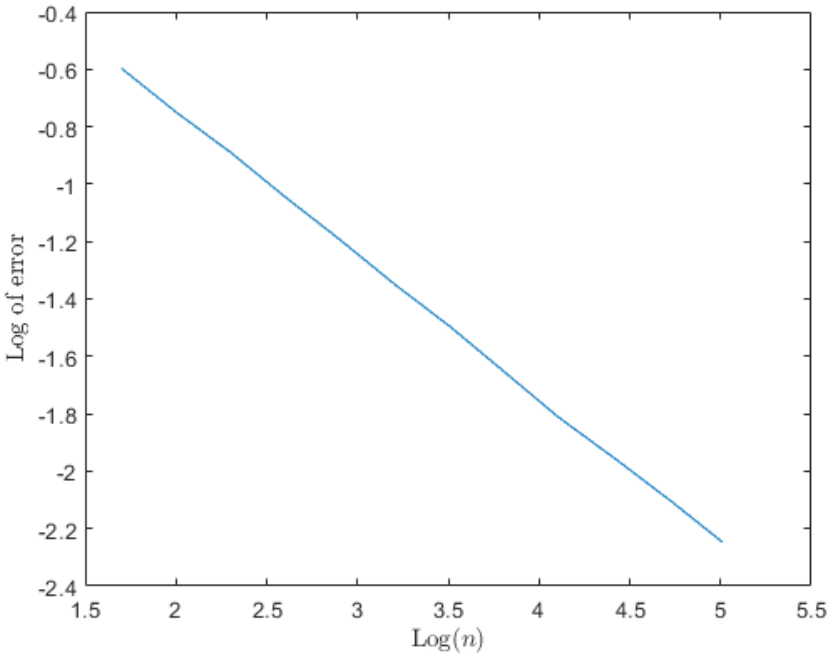

Figure 3.7: Plot of the random walk ℓ^2 error.

between the error of the optimal number of particles compared to the maximum number of particles. The random walk method Figure 3.7 shows a greater difference, but it is about half an order of magnitude. This demonstrates that the COMIC-generated optimal number of particles provides a strong approximate solution compared to that of the maximum number of particles. Even with a different metric, we were able to demonstrate that the COMIC provides a useful guide regarding the choice of the number of particles to utilize within a simulation. Regarding the comparison between numerical methods, we observe that the both of the simulations that run with the COMIC optimal number of particles (300 for the MTPT method and 5000 for the RWPT method) display similar errors, which are about $10^{-1.6}$. This is another indicator that the mass transfer method requires fewer particles to achieve results similar to that of the random walk method.

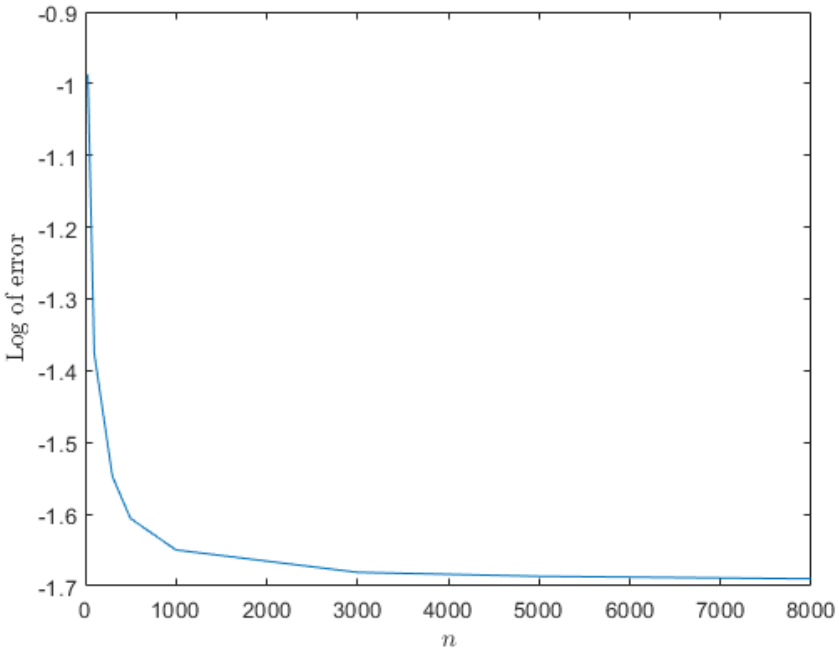

Figure 3.8: Plot of the mass transfer ℓ^2 error.

3.5 Summary

In this chapter, we demonstrated that particle methods which use the COMIC-generated optimal number of particles can be employed to accurately estimate parameters, i.e. velocity and diffusion coefficients, without incurring substantial error in comparison to those which involve greater numbers of particles. In addition, we verified that the COMIC produces a worthwhile reconstruction of the underlying concentration even with sets of noisy data that may better reflect real world data. To demonstrate this, we compared the AIC and COMIC models to the exact solution, and observed no significant differences. Finally, we used an ℓ^2 metric to quantitatively compare the error induced by the associated particle models.

CHAPTER 4
EXTENSIONS OF COMPUTATIONAL PROBLEM

In this chapter, we will examine the effects of constraints on real data to the COMIC. In particular, we will perform simulations with sparse, non-uniform data points and reformulate the information criterion as needed. Additionally, we will extend the definition of the COMIC to conform to the underlying assumption concerning the statistical errors between the data and the model.

4.1 Sparse Data Points

Another problem that may arise when collecting data is that the number of locations available for collection is limited. This may occur because the site is inaccessible due to contamination or from a lack of resources. The sparsity of data would likely require the method to utilize significantly more particles to better describe the underlying behavior of the system, which may affect the optimal number of particles predicted as in the previous section. In this section, we examine the effect of the lack of data on the COMIC. Also, we can see how well the previously predicted number of particles performs with this constraint. Instead of many data points (e.g., 30), we will only use 6 to 10 data points within the domain, and this will demonstrate both the degree to which parameter estimates are affected by less data and to what extent we need to increase the number of particles to achieve more precise estimates. Because we are relying on such a small collection of data, a small-data modification to the AIC is needed. This is often denoted as the AICc and defined by

$$\text{AICc} = \text{AIC} + \frac{2p^2 + 2p}{k - p - 1}, \tag{4.1}$$

where k is the number of data points, and p is the number of parameters in the model. This leads to the modification of the COMIC as

$$\text{COMICc} = \text{AICc} + \ln(n). \tag{4.2}$$

Such a correction is relevant when we are attempting to compare different models with various numbers of data points and parameters. However, in our simulations we do not change the number of data points or the number of parameters. Therefore, the latter term in the AICc will be identical amongst all models, and we will not include it in our calculation of the fitness metric. For the first simulation, we set $D = 1$, $v = 0$ and $T = 1$ and select 10 data points from the interval $[-5, 5]$ with uniform spacing. We perform the process of generating the COMIC-predicted optimal number of particles for both the random walk and mass transfer methods. For the former, the optimal number of particles increases, as seen in Figure 4.1, and the mass transfer simulation requires that the optimal number of particles range from 300 to 1000, as shown by Figure 4.2. Hence, we choose the new optimal number of particles for the random walk method to be 20000, but continue to use the same number of particles for the mass transfer method, namely 300.

Next, we will use the suggested optimal number of particles from the COMIC to estimate the value of D and v, in which we set $v = 1$ and the initial guess for both coefficients is 0.5. The estimated value of D in the MTPT is 1.0411, while v is estimated to be exactly its true value of 1. Additionally, the random walk simulation results are displayed in Figure 4.3.

Repeating this the simulation for 6 data points, the random walk requires significantly more particles. On the other hand, the mass transfer method displays similar behavior to other simulations - the COMIC optimal number of particles occurs at $n_{\text{COMIC}} = 300$ and the estimated values of D and v are 1.0321 and 1.0001, respectively.

4.2 Non-uniform Data

Data collection from a field site may not be constrained to a certain grid, meaning the data may not exist at exact gridpoints. Therefore, we will examine the effect of non-uniformly-spaced data on the COMIC and elucidate how this will affect parameter estimation. To generate non-uniform data, we randomly select data points within the domain of interest. More specifically, we perform simulations with 10 and 30 data points, with $D = 1$, $v = 1$, $x_0 = 0$ and $T = 1$. For RWPT method, we perform simulations with 5000 particles for 30

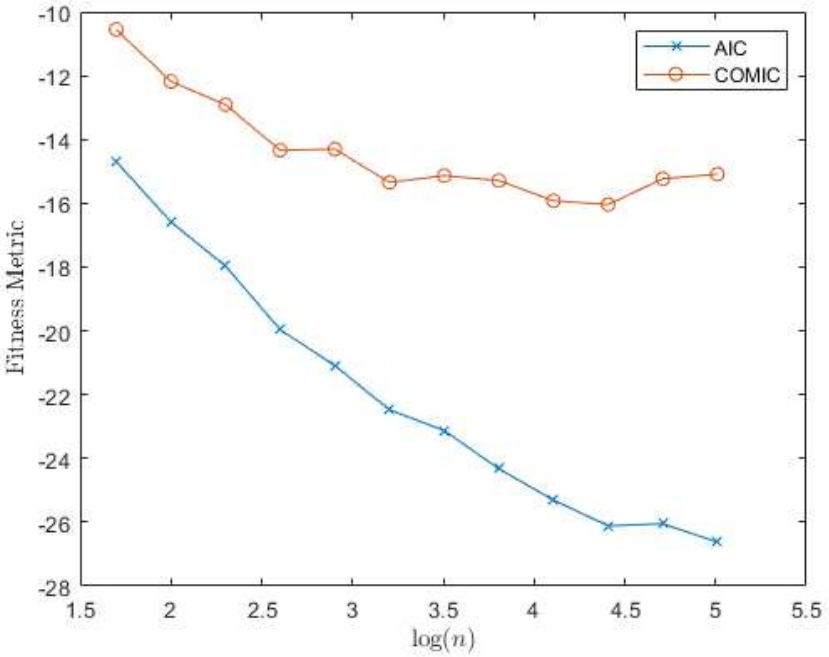

Figure 4.1: Plot of the random walk fitness metric for 10 data points.

data points and 20000 particles for 10 data points. Contrastingly, for the MTPT method both simulations utilize 300 particles. The initial guess for both parameters is 0.5 over all simulations. Due to the randomness in the spacing of the data, the results vary with each simulation, but in general the results are similar throughout all simulations. The results for random walk simulations are displayed in Figure 4.4 and Figure 4.5. For the mass transfer method, a simulation with 30 data points yields an estimate of 1.0385 for D and 0.9995 for v. Similarly, a simulation with 10 data points provides an estimate of 1.0380 for D and 0.9998 for v. From these simulations, we conclude that the COMIC provides a useful and informative guide for the choice of particle number even when the data is not uniformly spaced.

4.3 Non-Gaussian Error Processes

Instead of assuming the error of concentration at each grid point is normally-distributed with the same standard deviation. Chakraborty et al [5] showed, under no explicit as-

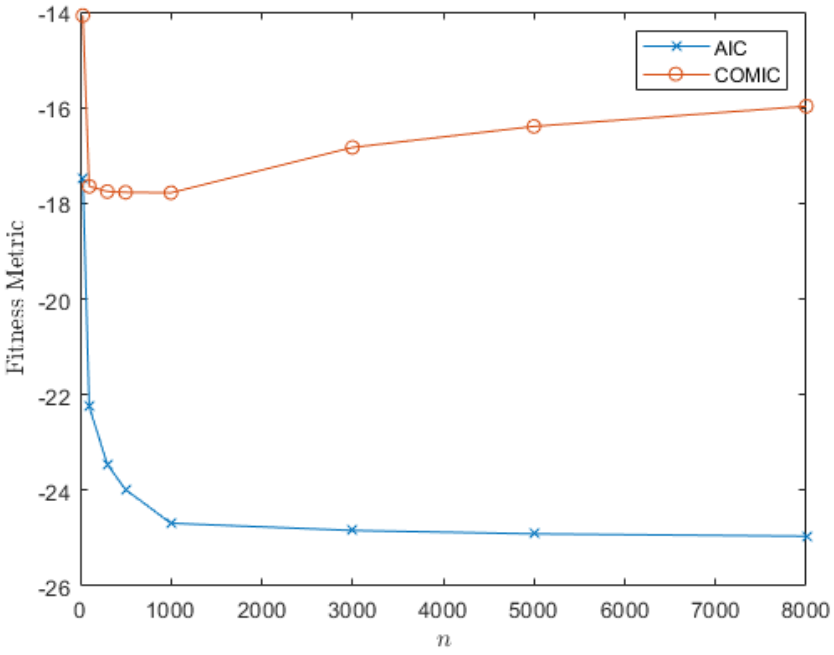

Figure 4.2: Plot of the mass transfer fitness metric for 10 data points.

sumptions on the error process, that (a) the concentration approximation generated by any particle method is proportionate to an asymptotically (as $n \to \infty$, $\Delta x \to 0$, and $\sqrt{n}\Delta x \to \infty$ where Δx is the bin size of the method) normal random variable and (b) the variance of the approximate concentration is actually proportionate to the concentration, i.e. $\sigma_i^2 = \alpha c(x_i, t)$ for some $\alpha \in \mathbb{R}$. In particular,

$$\alpha = \frac{m_{\text{total}}}{n\Delta x}. \tag{4.3}$$

This can be numerically verified by Figure 3.3 and Figure 3.4. For the mass transfer method, the bin size is the average particle spacing, i.e. the length of the domain divided by the number of particles, which implies $\alpha = 1/10$. We observe that the numerical results arising from a simulation with the COMIC-generated optimal number of particles or the maximum number of particles both give rise to greater errors near $x = 0$ and lesser discrepancies elsewhere. However, for the random walk method, the bin size is fixed, so α decreases as n increases, and this leads to lesser errors near $x = 0$ for the simulation that uses the maximum

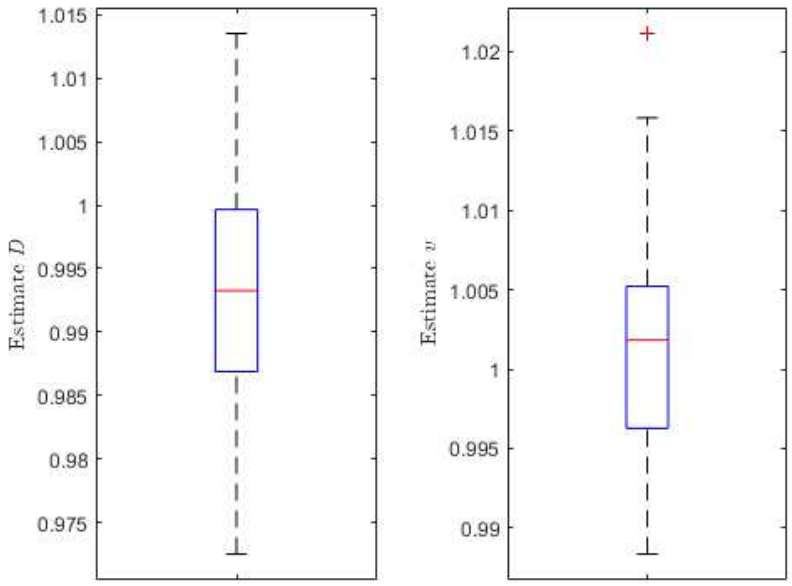

Figure 4.3: Plot of the random walk D and v estimate using 10 data points.

number of particles. With the above results, Chakraborty [5] also proposed an alternative information criterion for selecting a "best" model over all parameter choices with the very desirable property that the chosen parameter estimate $\hat{\theta}$ will serve as a consistent estimator for the true parameter values θ. More specifically, this criterion arises from an optimal fitting procedure that serves to minimize the weighted mean square error function

$$\mathcal{E}(\theta) = \frac{1}{k}\sum_{i=1}^{k} w_i|\hat{c}_i - c_n(x_i, T; \theta)|^2 \tag{4.4}$$

where θ is the vector of unknown model parameters (e.g., $\theta = [v, D]^T$), the minimization weights are

$$w_i = \frac{1}{m_{\text{total}}\hat{c}_i},$$

and the estimator $\hat{\theta}$ is given by

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \mathcal{E}(\theta).$$

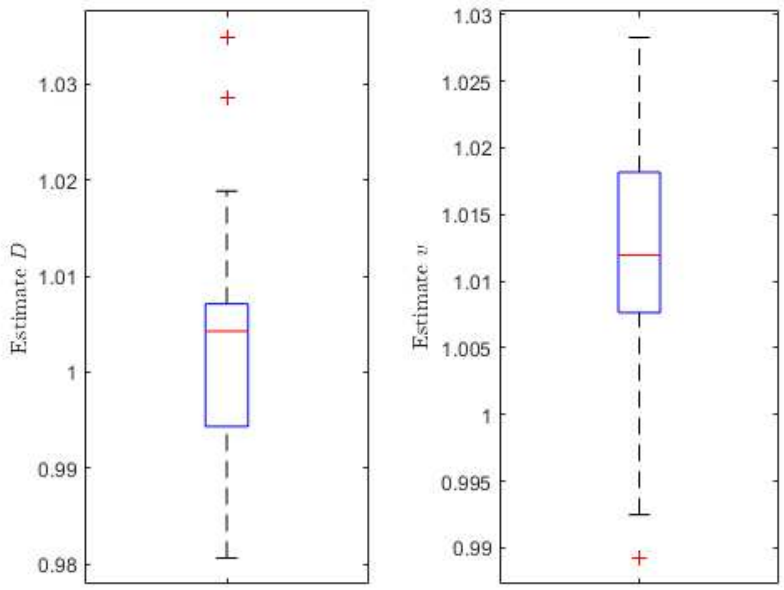

Figure 4.4: Plot of D and v estimate from 10 random data points.

For the ADE problem described in previous sections, we merely have $\theta = [v, D]^T$. Notice that in the case that the errors are normally-distributed, minimizing (4.4) is equivalent to maximizing the log-likelihood function (2.14) for a multivariate Gaussian distribution with $\hat{\sigma}_i^2 = m_{\text{total}}\hat{c}_i$ for $i = 1, \ldots, k$.

As for the AIC, this information criterion does not account for the additional information incurred by taking large numbers of particles, and hence we augment it to create a new computational information criterion. Hence, in this case we the define the COMIC by

$$\text{COMIC} = \ln(n) + 2\ln\left(\frac{1}{k}\sum_{i=1}^{k}\frac{1}{m_{\text{total}}\hat{c}_i}(\hat{c}_i - c_n(x_i, T))^2\right), \tag{4.5}$$

which is similar to (2.25), but due to the absence of normally-distributed errors, contains a different estimator for the variances $\hat{\sigma}_i^2$ for $i = 1, \ldots, k$.

Using this particular criterion, we perform simulations of the random walk and mass transfer methods to compute the value of $2\ln(\mathcal{E})$ and the COMIC. The formulation and implementation of these methods is analogous to that of the previous section with $k = 30$

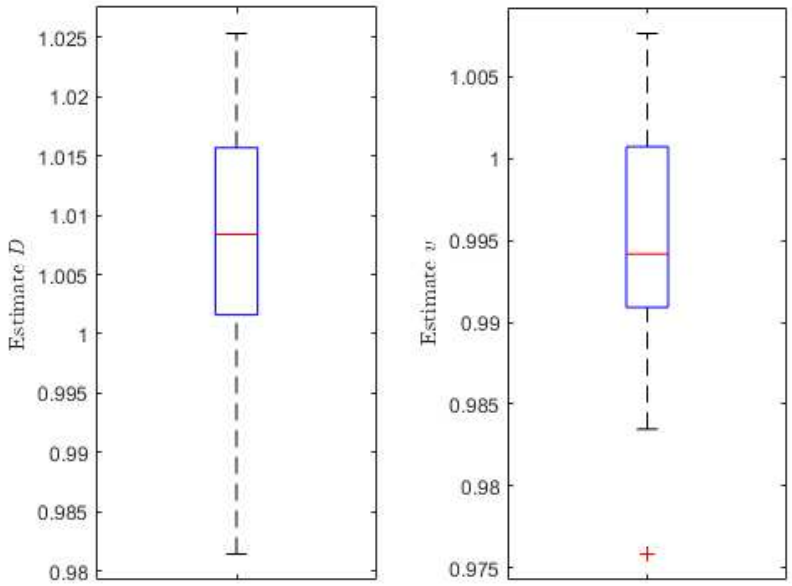

Figure 4.5: Plot of D and v estimate from 30 random data points.

randomly-spaced data points. From Figure 4.7, the optimal number of particles for the MTPT method is 300 particles, while for the RWPT method, it ranges between 3200 and 6400 particles, as shown in Figure 4.6. Similar to the previous section, we take the midpoint of these (3200 and 6400) and assume the optimal number of particles is about 5000. Notice that this predicted value is the same as that stemming from the single Gaussian error simulations. Then, we use these optimal number of particles to perform parameter estimation as in the previous section, which provides MTPT estimates of $D = 1.0253$ and $v = 1.0020$. Because the data are random, the results may vary; however, multiple runs with different data display similar results - see Figure 4.8 for RWPT results. The maximal absolute error in the estimate of D is about 2.5% and for v it is around 4%. These results are comparable to the parameter estimation performed by minimizing the maximum likelihood function. Therefore, the COMIC demonstrates consistency among different error assumptions and estimators.

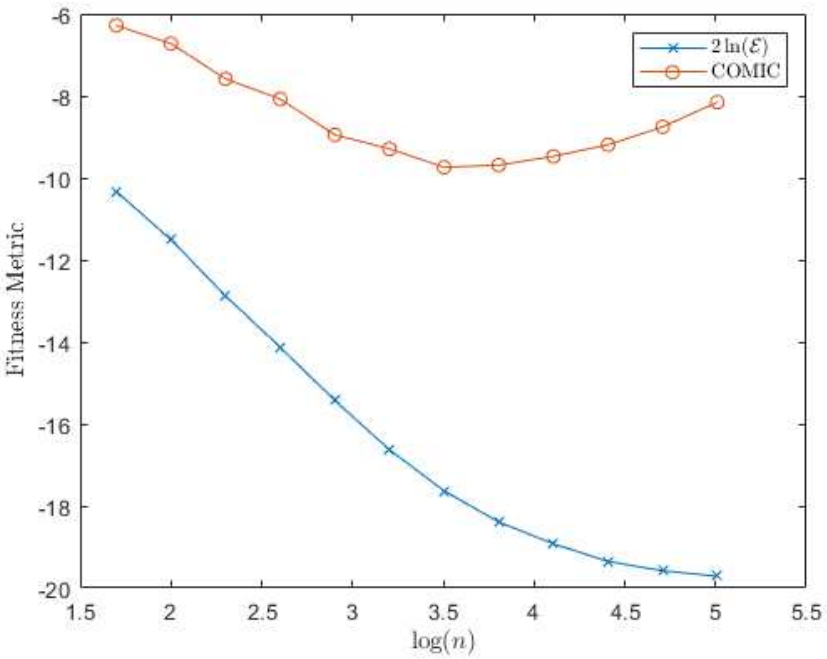

Figure 4.6: Plot of the random walk fitness metric for non-Gaussian error processes.

4.4 Summary

In this chapter, we extended the COMIC to the case of sparse data points, which is similar to the extension of the AIC to the AICc. Under this condition, the COMIC optimal number of particles for the mass transfer does not vary, but the random walk method requires greater particle numbers as the number of data points decreases. In addition, we performed parameter estimate simulations with data that was not uniformly-spaced, and the results are similar to the case of uniform data. Lastly, we explored non-Gaussian error distributions, and hence the use of a consistent estimator, that is not an MLE, to define the COMIC.

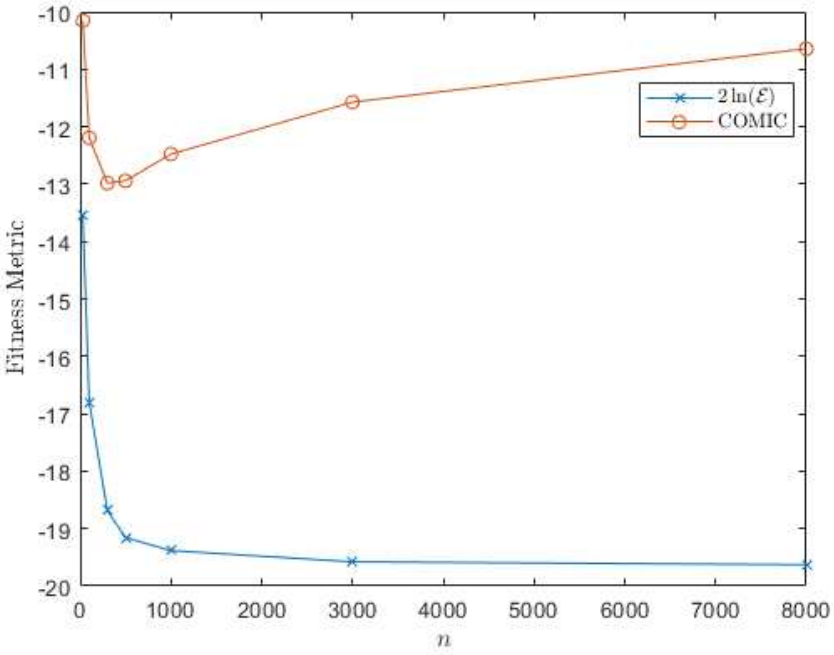

Figure 4.7: Plot of the mass transfer fitness metric for non-Gaussian error processes.

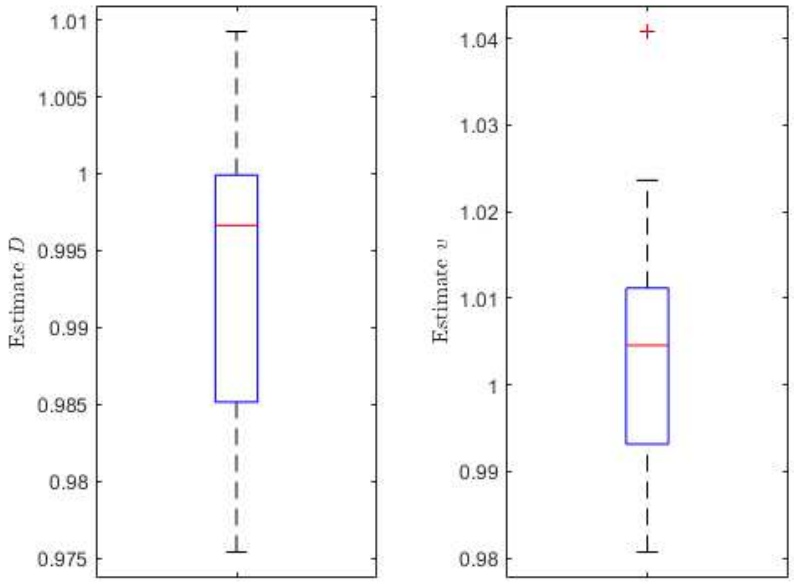

Figure 4.8: Plot of the random walk D and v estimate for non-Gaussian error processes.

REFERENCES CITED

[1] G. Pavliotis and A. Stuart. *Multiscale Methods: Averaging and Homogenization*. Springer Science & Business Media, 2008.

[2] P. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1992.

[3] M. Hill and C. Tiedeman. *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*. John Wiley & Sons, 2007.

[4] P. Brockwell and R. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics, 3d edition, 2016.

[5] P. Chakraborty, M. Meerschaert, and C. Lim. Parameter estimation for fractional transport: A particle-tracking approach. *Water Resources Research*, 45(10):W10415, 2009.

[6] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19(6):716–723, 1974.

[7] D. Benson, S. Pankavich, M. Schmidt, and G. Sole-Mari. Entropy: (1) the former trouble with particle-tracking simulation and (2) a measure of computational information penalty. *Advances in Water Resources*, 137:103509, 2020.

[8] D. Benson and D. Bolster. Arbitrarily complex chemical reactions on particles. *Water Resources Research*, 52(11):9190–9200, 2016.

[9] Michael J. Schmidt, Stephen D. Pankavich, and David A. Benson. On the accuracy of simulating mixing by random-walk particle-based mass-transfer algorithms. *Advances in Water Resources*, 117:115–119, 2018.

[10] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Springer Series in Statistics, Perspectives in Statistics*, pages 610–624, 1992.

[11] P. Kitanidis. The concept of the Dilution Index. *Water Resources Research*, 30(7): 2011–2026, 1994.

[12] S. Konishi and G. Kitagawa. *Information Criteria and Statistical Modeling*. Springer Series in Statistics. Springer, New York, NY, 2008.

[13] S. Kullback. *Information Theory and Statistics*. Dover Publications, 1968.

[14] M. Rahbaralam, D. Fernàndez-Garcia, and X. Sanchez-Vila. Do we really need a large number of particles to simulate bimolecular reactive transport with random walk methods? a kernel density estimation approach. *Journal of Computational Physics*, 303: 95–104, 2015.