Inference of Population History in an Isolation-with-Migration Model

Challenges in Bioinformatics
Quarter 1 · October 2013

Bioinformatics Research Centre
Computer Science Department
Aarhus University

By

成玉

Jade Y. Cheng

Supervisor: Thomas Mailund, PhD
Professor: Jotun Hein, PhD

Abstract

Traces and hints of evolution history are embedded in the genomic data found in present-day populations. Facilitated by next generation sequencing technology, an overwhelming volume of genomic data has become accessible. In this study we strive to recover the historical process of population divergence and gene flow in two isolated sub-populations.

We describe the two sub-populations using an Isolation-with-Migration (IM) model. We formulate an analytical process to infer the population parameters in the IM model using the Maximum Likelihood Estimation (MLE) with the matrix exponential framework. We consider pairwise samples with different initial configurations from the IM model, and we use their divergence times to perform population history inference.

The inference procedure was developed as software and evaluated with a series of experiments. In solving a one-epoch system parametrized with coalescent rates and migration rates, the presented mechanism achieved a high accuracy. It fully recovered the population history in an one-epoch system. In a multi-epoch system, however, the procedure failed to make accurate inferences under the given conditions. The software was tuned to achieve its best performance. The affecting factors are presented as a sequences of observations.

Finally, we discuss the two aspects of the future directions of this project, the main framework of the inference procedure and improvements on the implementation and execution. For the former, we consider modifications of the presented inference framework. Mostly, we raise questions regarding the true genealogy, different sampling techniques, and formulating parameters as functions of time. For the latter, we consider the time complexity of the current implementation and the possibility of parallelization.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Population History Inference

Demography influences the pattern of genetic variation in a population, thus genomic data of multiple individuals sampled from one or more present-day populations contains valuable information about the past demographic history [1]. In the study of speciation, researchers often inquire about the extent that populations have exchanged genes since those populations began to diverge. Answers to questions about historical divergence and gene flow potentially lie in patterns of genetic variation that are found in present-day populations [2].

1.2 Isolation-with-Migration Model

An IM model incorporates both population separation and migration [3]. Under an IM model the genealogies include not only some fixed number of coalescent events and speciation events, but they also include any possible number of migration events. We employ a two-population IM model. It is parameterized by the coalescent rates of lineages and the rates of migration between populations. Figure 1 is an illustration of such a model with a single epoch.

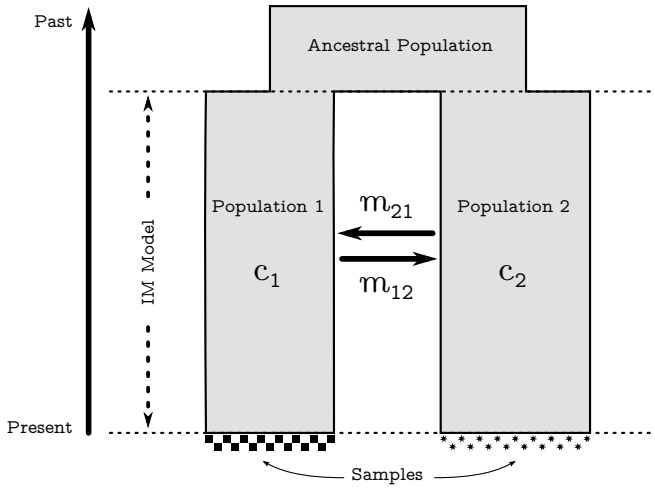

Figure 1: Isolation-with-Migration Model

1.3 Pairwise Samples

We examine the simple case in which only a pair of genes are sampled from two populations, shown in Figure 2. When considering only pairs of samples, the likelihood of a model will depend on only the divergence time at each locus. We explicitly consider the coalescent process for pairs of samples and derive the exact transition probabilities from this model.

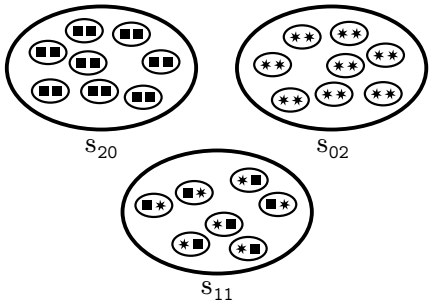

Figure 2: Pair-Wise Sample Configurations

2 Methods

To bridge the gap between population history and current genetic data, population geneticists can make use of a gene genealogy, G, a bifurcating tree that represents the history of ancestry of sampled gene copies. The probability of a particular value of G can be calculated for a particular parameter set using coalescent models. Then given a particular genealogy, genetic variation can be examined using a mutation model that is appropriate for the kind of data being used. Finally, by considering multiple values of G, the connection can be made between the population evolution history and the data [4, 2].

For each locus, the data D consists of two samples, and therefore the probability of the data depends on only the time to the most recent common ancestor (MRCA); i.e., the divergence time at this locus. We can write the likelihood for a single locus as:

$$L(\Theta \mid D) = P(D \mid \Theta) = \int_t^\infty [P(D \mid t) \times P(t \mid \Theta)]\, dt$$

The product of likelihood over all loci can be written as:

$$\prod_D [P(D \mid \Theta)] = \prod_D \left\{ \int_t^\infty [P(D \mid t) \times P(t \mid \Theta)]\, dt \right\}$$

The log likelihood preserves the maximum and minimum features of the likelihood function, and it is computationally easier to handle, so we work with the following log likelihood directly:

$$\sum_D \log [P(D \mid \Theta)] = \sum_D \log \left\{ \int_t^\infty [P(D \mid t) \times P(t \mid \Theta)]\, dt \right\}$$

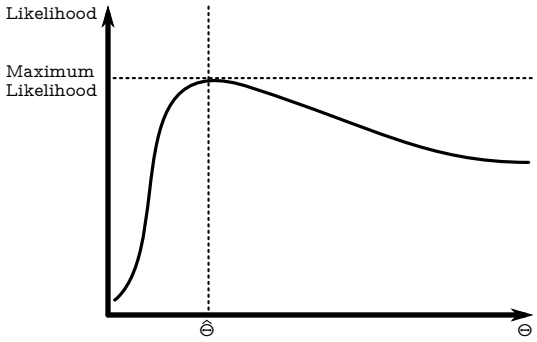

Figure 3: Maximum Likelihood Estimate

In this study, we obtain $P(t \mid \Theta)$ using the matrix exponential framework, and we obtain $P(D \mid t)$ with simulated TMRCAs between pairwise samples. According to the nature of MLE, if the model is true, with a increased sample size, the parameter estimate should approach the true parameters.

$$\hat{\Theta} \quad \rightarrow \quad \Theta_{true}$$

2.1 States and State Transitions

We consider events that occur after the speciation time, T. For two samples the system is in one of the following four states: S_{20}, where both samples are in population 1; S_{02}, where both samples are in population 2; S_{11}, where one sample is in population 1 and the other is in population 2; and C, where the two samples have coalesced and the single locus is in either population 1 or population 2.

2.1.1 Migration Events

A migration event results in a specific switch from one state to another, illustrated in Figure 4. State transitions caused by migration events do not involve the coalesced state C because migration events themselves do not cause two samples to coalesce.

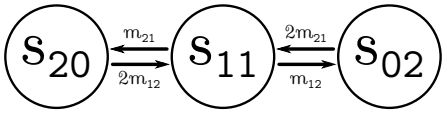

Figure 4: State Transitions Caused by Migration Events

2.1.2 Coalescence Events

A coalescent event also results in a specific switch from one state to another, illustrated in Figure 5. This state transition can originate from only two states, S_{20} and S_{02}, because the pairwise samples need to be physically present in the same population before a coalescent event can occur. Once two samples reach the coalescent state, they do not escape from this state.

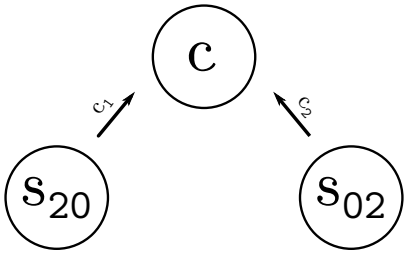

Figure 5: State Transitions Caused by Coalescent Events

2.1.3 State Transitions

Combining the state transitions caused by migration events and state transitions caused by coalescent events, we reach the overall state transition relationship, illustrated in Figure 6.

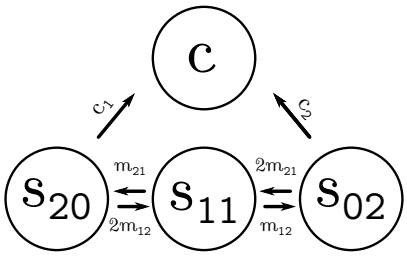

Figure 6: State Transitions

2.2 Matrix Exponential Framework

Let Q denote the rate matrix of the continuous time Markov chain (CTMC). According to CTMC theory, to integrate over all paths leading from one state to another, we can simply perform the matrix exponential operation on Q. The probability of being in state s at time t and state s' at time t' is given by $P_{s,s'}^{t'-t}$ where $P_{s,s'}^{t'-t} = \exp\left(Q \cdot (t'-t)\right)$. This framework has a clear advantage over explicitly integrating over all possible sample paths in the system [6].

2.2.1 The Rate Matrix

We form the rate matrix Q for this CTMC according to the state transition diagram discussed in Section 2.1. The diagonal values are assigned so the sum over all entries of each row is zero.

$$
Q = \begin{array}{c} \\ S_{11} \\ S_{20} \\ S_{02} \\ C \end{array}
\begin{array}{c}
\begin{array}{cccc} S_{11} & S_{20} & S_{02} & C \end{array} \\
\begin{bmatrix}
-(m_{12}+m_{21}) & m_{21} & m_{12} & 0 \\
2 \cdot m_{12} & -(2 \cdot m_{12} + c_1) & 0 & c_1 \\
2 \cdot m_{21} & 0 & -(2 \cdot m_{21} + c_2) & c_2 \\
0 & 0 & 0 & 0
\end{bmatrix}
\end{array}
$$

2.2.2 From Q to $P\,(T < t)$

Computing the matrix exponential is an expensive operation, and it is implemented with a variety of approximation algorithms [8]. For example, truncating from the Taylor series is one such approximation algorithm.

$$\begin{aligned} P &= MatrixExp\,(Q \times t) \\ &= \sum_{i=0}^{\infty} \frac{(t \times Q)^i}{i!} \\ &= I + t \times Q + \frac{(t \times Q)^2}{2!} + \cdots + \frac{(t \times Q)^k}{k!} + \cdots \end{aligned}$$

Each P matrix obtained from this calculation reveals the probabilities of a pairwise sample starting from certain states and arriving at certain other states by time t. For example, in the expression below, if we were interested in the probability of arriving at state C after time t for a pairwise sample that started from state S_{11}, we would look at the entry labeled with $f_{S_{11}\to C}$, which corresponds to the first row and last column in this P matrix.

$$P\,(T < t) = \begin{array}{c} \\ S_{11} \\ S_{20} \\ S_{02} \\ C \end{array} \begin{array}{c} \begin{array}{cccc} S_{11} & S_{20} & S_{02} & C \end{array} \\ \begin{bmatrix} 1 - \sum & f_{S_{11}\to S_{20}} & f_{S_{11}\to S_{20}} & f_{S_{11}\to C} \\ f_{S_{20}\to S_{11}} & 1 - \sum & f_{S_{20}\to S_{02}} & f_{S_{20}\to C} \\ f_{S_{02}\to S_{11}} & f_{S_{02}\to S_{20}} & 1 - \sum & f_{S_{02}\to C} \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

2.2.3 From $P\,(T < t)$ to $f\,(T = t)$

In a continuous case, to convert the probability described as a CDF into a PDF, we take the derivative. In this study, since we are taking discrete time steps going backwards in time, to retrieve the probability density from the expression discussed in Section 2.2.2, we would simply perform the following subtraction using the P matrices calculated for each discrete time step.

$$f\,(t_{i-1} < T < t_i) = P\,(T < t_i) - P\,(T < t_{i-1})$$

For this study, we are interested in the state transitions between all initial states and the coalesced state. This provides us information about the probabilities of pairwise samples coalescing at certain time periods. With a small abuse of notation, we can write the probability density functions as below.

$$f_{S_{20}\to C}\,(T = t) = \begin{cases} P_{S_{20}\to C}\,(T < t_1) - P_{S_{20}\to C}\,(T < t_0) \\ \vdots \\ P_{S_{20}\to C}\,(T < t_k) - P_{S_{20}\to C}\,(T < t_{k-1}) \\ \vdots \\ 1 - P_{S_{20}\to C}\,(T < t_n) \end{cases}$$

$$
f_{S_{11}\to C}(T=t) = \begin{cases} P_{S_{11}\to C}(T<t_1) - P_{S_{11}\to C}(T<t_0) \\ \vdots \\ P_{S_{11}\to C}(T<t_k) - P_{S_{11}\to C}(T<t_{k-1}) \\ \vdots \\ 1 - P_{S_{11}\to C}(T<t_n) \end{cases}
$$

$$
f_{S_{02}\to C}(T=t) = \begin{cases} P_{S_{02}\to C}(T<t_1) - P_{S_{02}\to C}(T<t_0) \\ \vdots \\ P_{S_{02}\to C}(T<t_k) - P_{S_{02}\to C}(T<t_{k-1}) \\ \vdots \\ 1 - P_{S_{02}\to C}(T<t_n) \end{cases}
$$

2.3 Simulated TMCRAs

To perform experiments and evaluate the inference capability of our model, we need to simulate sample data. Specifically, we need to simulate the TMCRAs for pairwise samples drawn from three initial configurations. These configurations correspond to the three pairwise states discussed in Section 2.1. Figure 7 illustrates these configurations.

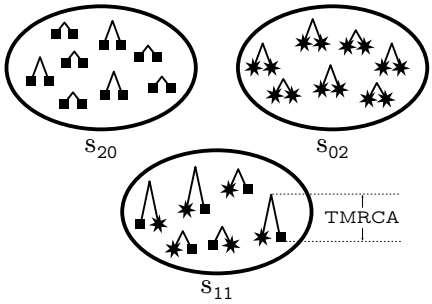

Figure 7: TMCRA Simulation

2.3.1 The Sample Maker Utility

In this study, we used the `ms` program from Richard Hudson. Program `ms` generates samples under neutral models. It is widely used for coalescence simulations. Program `ms` is capable for a wide range of types of coalescence simulations, such as simulating population bottlenecks, growth, recombinations, etc [7]. The following command is an example use case employed in this study.

```
$ ./ms 2 500 -T -I 2 2 0 -m 1 2 0.2 -m 2 1 0.4 -n 1 1.0 -n 2 0.5
```

In this execution, we generate 500 pairwise samples expressed as the tree format. Samples are simulated in a two-population isolation model with specified migration rates and populations sizes. Specifically, Table 1 details the command-line arguments used above.

Argument	Description
2	to generate pairwise samples
500	number of samples to be simulated
-T	output as the tree format; i.e., for pairwise samples output a scalar, the divergence time
-I 2	isolation model with two sub-populations
2 0	pairwise sample is simulated from population 1; it could also be 1 1 or 0 2
-m 1 2 0.2	the scaled migration rate from population 1 to population 2 is 0.2
-m 2 1 0.4	the scaled migration rate from population 2 to population 1 is 0.4
-n 1 1.0	the scaled population size in population 1 is 1.0
-n 2 0.5	the scaled population size in population 2 is 0.5

Table 1: Arguments used in the example execution of the `ms` Program

Notice the coalescent rates are not directly used as arguments to the `ms` program during data simulation. The scaled population sizes are used instead. The coalescent rates correspond to the reverse of the scaled population sizes; i.e. $c = 1/n$.

2.3.2 Distribution of Simulated TMRCAs

Pairwise samples for each configuration are simulated as a forest of trees. Each tree contains two leaves. The only information recorded for these trees is their tree heights, which corresponds to the TMRCAs for the pairwise samples represented as the tree leaves. For each configuration, histograms were created to visualize the distributions of the TMRCAs under a certain parameter set. Overall these distributions fit with our expectations for the given parameter sets.

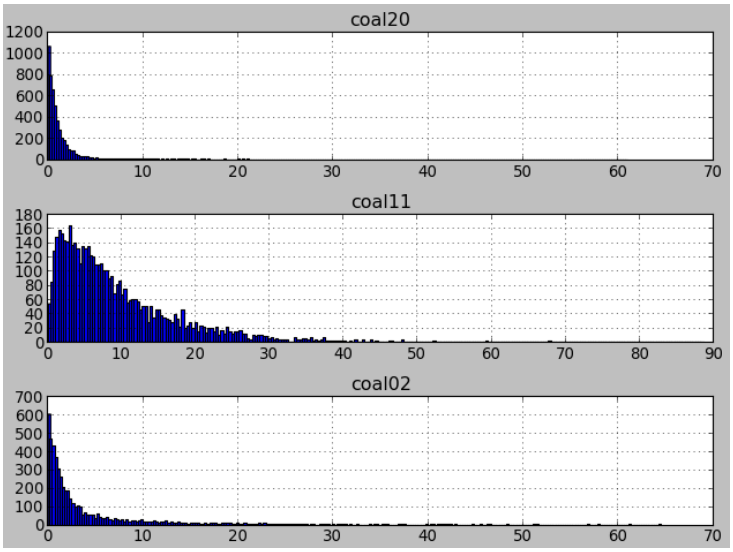

Figure 8: Simulated TMCRA Data – Moderate Population Sizes

The histogram above was created from samples generated with the `ms` program using the following parameters: $c_1 = 1$, $c_2 = 0.5$, $m_{12} = 0.1$, and $m_{21} = 0.2$.

The simulated TMCRAs for configuration S_{20} and S_{02} roughly form exponential distributions. Under the standard population model (fixed population size and no migrations), the TMCRAs should follow the exponential distribution. In this simulation, however, we have a small amount of migration, so the TMCRAs for S_{20} and S_{02} are not strictly exponentially distributed. But they are close because the coalescent events dominate the migration events.

The distribution of the simulated TMCRAs for configuration S_{11} is significantly different from the distributions observed in S_{20} and S_{02}. Migration is crucially involved in determining the TMCRAs when the pairwise samples originate from different sub-populations. In other words, the pairwise samples need to first make their way into one single population through migration before they have a chance to coalesce. This also involves migrating back-and-forth multiple times between populations before coalescing.

Finally, it is clear there is a significant difference in time to coalesce between pairwise samples from the same population and pairwise samples from different populations. The interpretation of this phenomenon also lies in the involvement of the migration events.

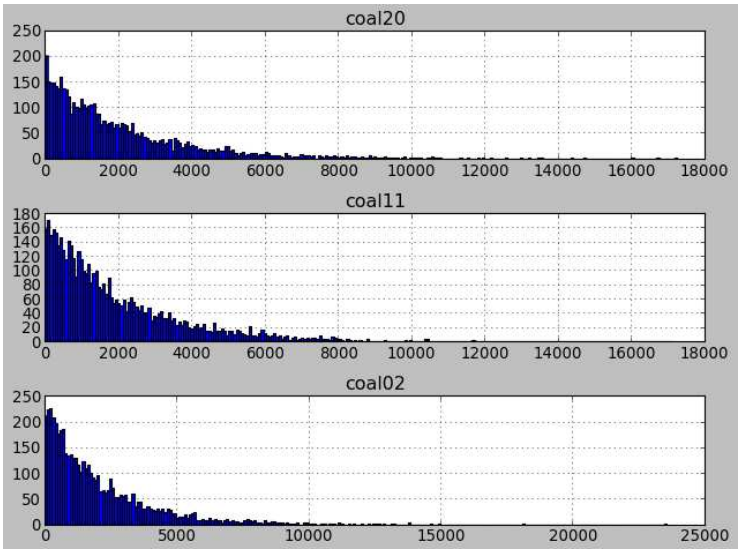

Figure 9: Simulated TMCRA Data – Large Population Sizes

The histogram in Figure 9 was created from samples generated with the `ms` program using the following parameters: $c_1 = 0.0001$, $c_2 = 0.0005$, $m_{12} = 0.1$, and $m_{21} = 0.2$. The TMCRA distributions observed from samples simulated with this set of parameters noticeably differ from what we saw in Figure 8. The three-sample configurations tend to have the same distribution. This is caused by significantly reduced coalescent rates. With small coalescent rates (large population sizes) the coalescent events do not dominate migration events. When migration thrives, sub-populations can be approximated as a single population. Hence the TMCRA distribution from pairwise samples originated from different populations resembles the characteristics of a single population.

3 Implementation and Experiments

The `Pairwise − IM` program was developed and tested using Python 2.7.3 on 64-bit GNU/Linux. It implements the parameter inference mechanism for an IM system utilizing MLE with the matrix exponential framework. The application consists of three main Python modules: (i) the algorithms module, which implements MLE and the matrix exponential framework; (ii) the sample maker module, which communicates with the `ms` program; and (iii) the testing module, which grows as more analysis and experiments are designed.

In this section, we first describe the possible issues with global optima and local optima in MLE. We then show experiments demonstrating informative inferences and uninformative inferences. Finally, we show the results of applying this method to perform population history inference in a single-epoch IM system as well as a multi-epoch IM system.

3.1 Global Optima vs. Local Optima

A local optima of an optimization problem is a solution that is optimal, either maximal or minimal, within a neighboring set of candidate solutions. A global optima is the optimal solution among all possible solutions, not just those in a particular neighborhood of values. Global optima are always local optima while local optima may not be global optima. To perform a preliminary check on the presented process regarding this issue, example executions are viewed from different perspectives, shown in Figure 10.

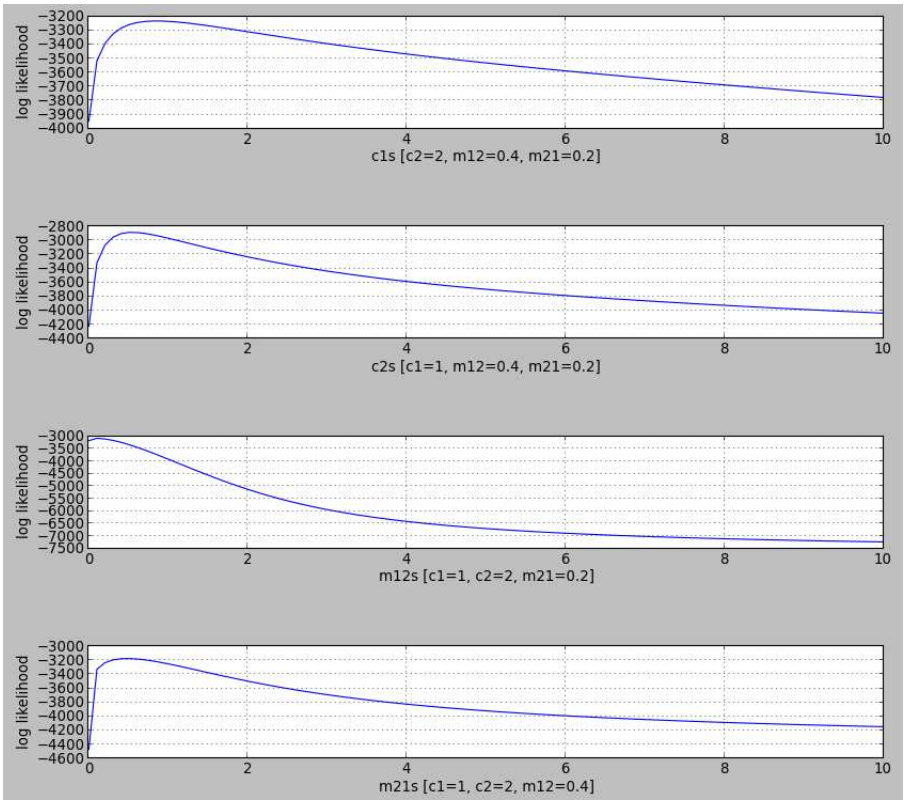

Figure 10: Global Optima vs. Local Optima

For each chart, one parameter was inferred while other parameters were kept at the values used for data simulation. On the Y axis, each plot shows the log likelihood values for a single-epoch system with a single varying parameter. The X axis holds the varying parameter. We see only a single peak from all four different angles. This demonstrates the MLE methodology with the matrix exponential framework has a potential to work well with population history inference in a single-epoch system.

3.2 Informative Inference vs. Uninformative Inference

In each inference execution, parameters are solved, but the quality of inference varies as the informative level of each execution on a particular data set varies.

3.2.1 Log Likelihood View

Figure 11 and 12 demonstrate an informative inference and an uninformative inference, respectively. In each chart, the Z axis holds the log likelihood values. The X and Y axes hold the varying parameters that the system is trying to solve. The rest of the parameters were kept fixed at the values used for data simulations. We see Figure 11 has one single peak. The projection of this log likelihood value onto the X, Y plane gives the inferred values for the parameters being solved. We would expect a good recovery of these parameters since the log likelihood has a very clear peak.

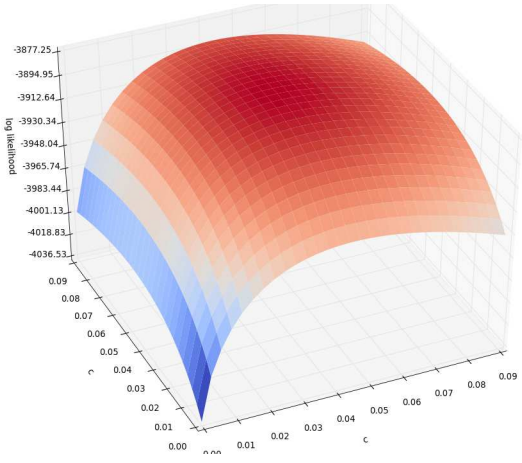

Figure 11: An Example of Informative Inference

We see in Figure 12, the log likelihood values from all parameter sets roughly form a plane without a clear peak value. The system would still attempt to reach a high point among all of these log likelihood values. We would not, however, expect a good recovery of the parameters from this execution. The log likelihood values do not seem to contain a sufficient amount of information to describe these parameters. In other words, varying these parameters does not appear to have a significant enough effect on the likelihood function.

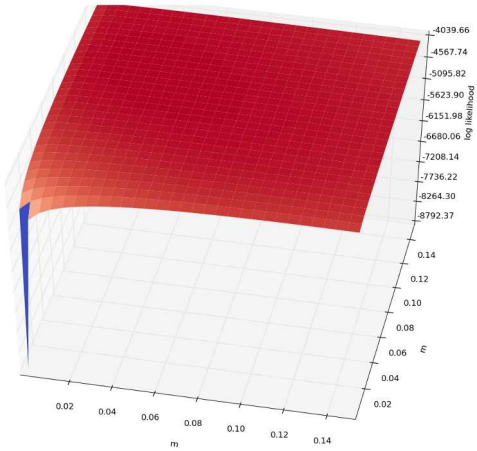

Figure 12: An Example of Uninformative Inference

Notice that a single view of plane-shaped log likelihood values is not enough to conclude that the inference is uninformative. The actual differences of these log likelihood values matter. In other words, if we zoom in and increase the resolution on the Z axis, the shape might become more like Figure 11 rather than Figure 12.

3.2.2 Parameter Inference View

Figure 13, and 14 demonstrate a series of informative inferences in a one-epoch system.

There were 15 and 50 linearly increasing c_1 values used as the parameters to solve. The rest of the parameters were fixed at the values used during data simulation, specifically, $c_2 = 0.5$, $m_{12} = 0.05$, and $m_{21} = 0.1$. Each box represents a set of 20 independent executions on a given c_1. The red dash inside the box indicates the mean value of these 20 solved c_1 values. The blue outline of the box indicates the range (0.25 mean and 0.75 mean) of these 20 solved c_1 values. The dashes outside the blue boxes indicate extreme values of the solved c_1 values. For each independent execution, three sets of pairwise samples were generated. Each configuration contains 500 sample TMCRAs. We see the presented system solves the coalescent rate parameters in a one-epoch system with good accuracy.

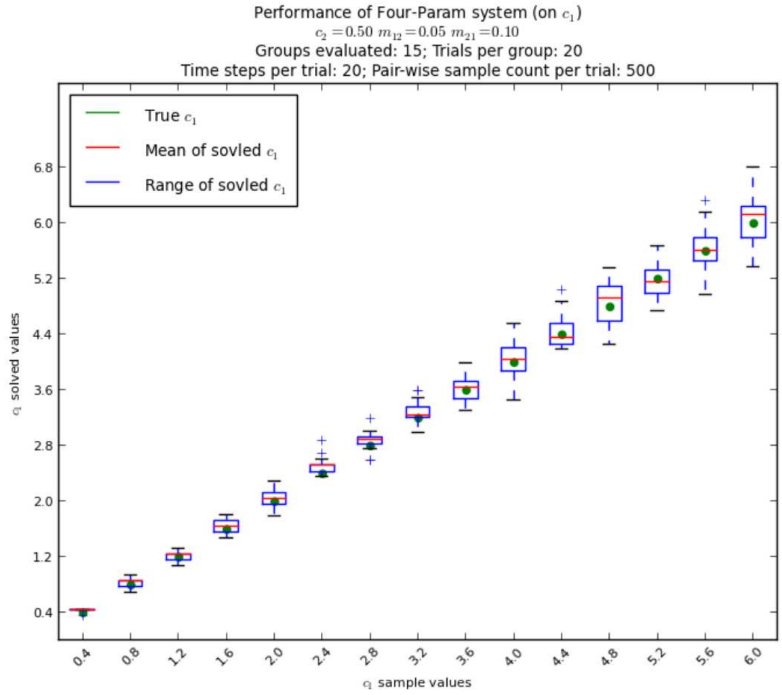

Figure 13: Inference Produced by Informative Datasets – 15 Data Groups

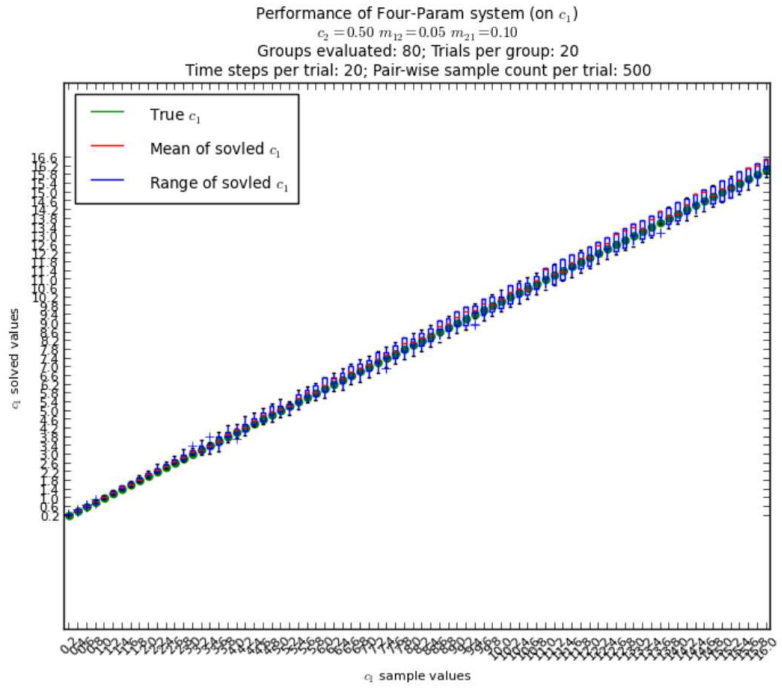

Figure 14: Inference Produced by Informative Datasets – 50 Data Groups

Figure 15 and 16 demonstrate a series of less informative inferences in a one-epoch system.

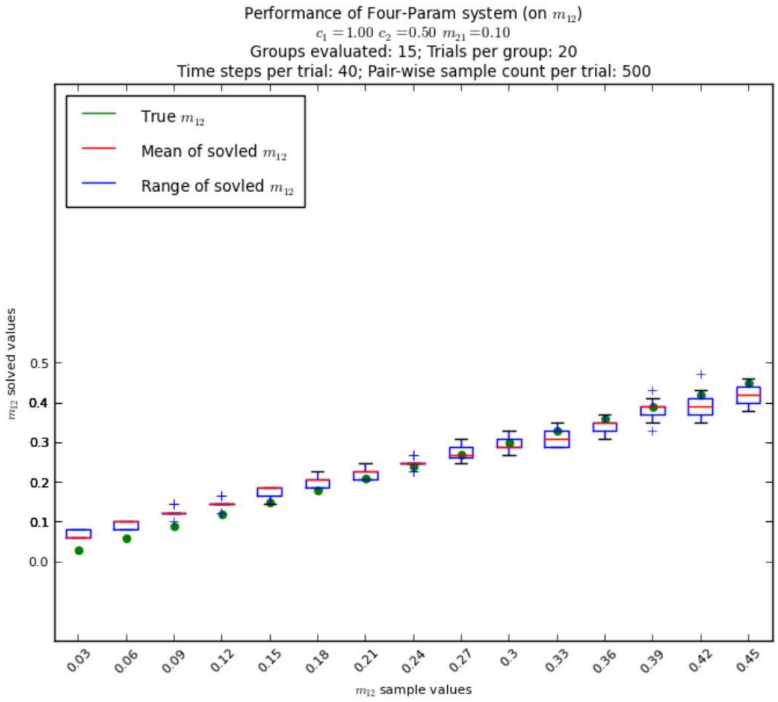

Figure 15: Inference Produced by Less Informative Datasets – 15 Data Groups

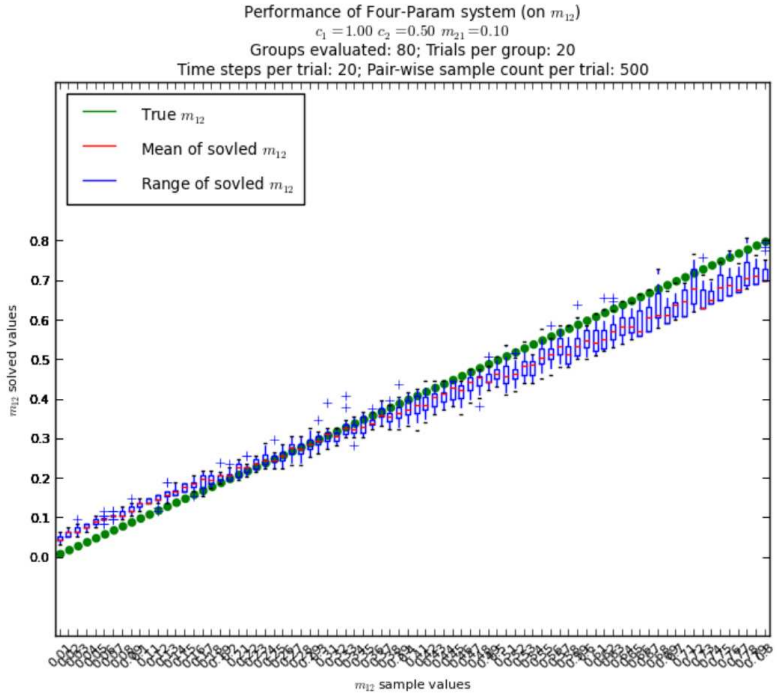

Figure 16: Inference Produced by Less Informative Datasets – 50 Data Groups

There were 15 and 50 linearly increasing m_{12} values used as the parameters to solved. The rest of the parameters were fixed at the values used during data simulation, specifically, $c_1 = 1.0$, $c_2 = 0.5$, and

$m_{21} = 0.1$. The remaining portion of this experiment is identical to the one shown in Figure 13 and 14.

We see the presented system does not solve the migration rate parameters in a one-epoch system well.This inaccuracy could be the result of several different factors. For instance, the sample data set could be too small. Or, the time steps could be too big and the migration events were not captured. Refer to Section 4.1 and 4.2 for further investigation of this issue.

On the other hand, we notice a significant and consistent bias. This phenomenon is especially visible on Figure 16. For smaller values, overestimates were produced, while for larger values, underestimates were produced. We also see the solve parameters follow the general trend of the true parameters, i.e., solved parameters increase as the the true parameters increase. This implies the sample data size might not be adequate. Larger volume sizes were tested and will be discussed in Section 4.2.

3.3 Population History Inference

To evaluate the inference mechanism in a more systematic manner, we simulated the parameters to solved in a one-epoch system with following normal distributions.

$$\begin{aligned} c_1 &\sim \mathcal{N}(2.0,\ 0.5) \\ c_2 &\sim \mathcal{N}(2.0,\ 0.5) \\ m_{12} &\sim \mathcal{N}(1.0,\ 0.2) \\ m_{21} &\sim \mathcal{N}(1.0,\ 0.2) \end{aligned}$$

To visualize the performance, we designed three types of scatter charts. In Type I, pairs of parameters used to generate the sample data were plotted. In Type II, the differences between solved and sampled values for pairs of parameters were plotted. In Type III, the sampled and solved values for one parameter were plotted.

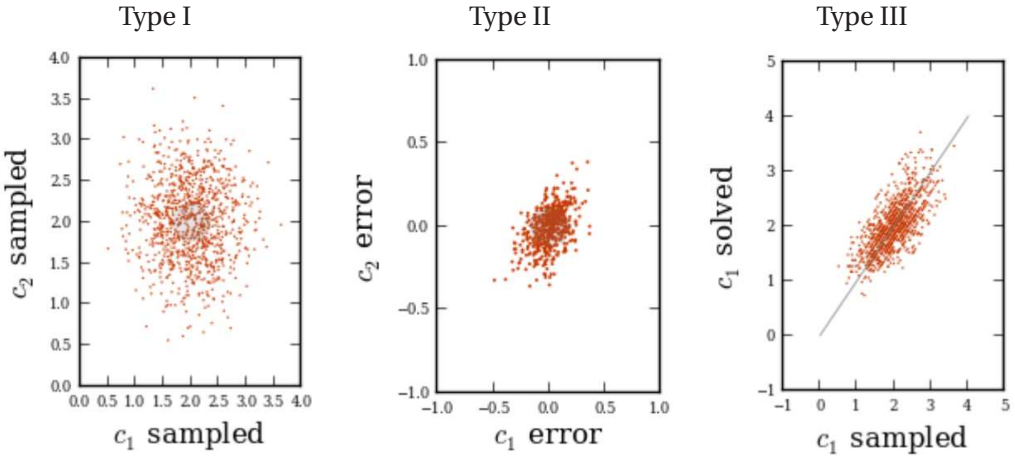

Table 2: Three Types of Scatter Plots

In Type I, we expect to see a circular Gaussian cloud centered at the mean values of the plotted parameters. In Type II, we expect to see all data points centered at the origin, preferably with a small diameter. In Type III, we expect to see all data points aligned along $y = x$ with the highest density at the mean values of the plotted parameters.

3.3.1 Population History Inference on a Single-Epoch System

In a single-epoch system, the inference process solves four parameters: c_1, c_2, m_{12}, and m_{21}. The overall achievement of the presented mechanism is promising. Different selections of various constants used in the data simulation affect the inference results. The following constants were used in the experiment presented in Figure 17.

```
# of parameter combinations tested (# of red dots in each plot)    500
For each dot:  # of data points tested in a range                  200
For each dot:  # of steps in each epoch                            20
For each dot:  # of samples for each 11, 20, and 02 configuration  500
```

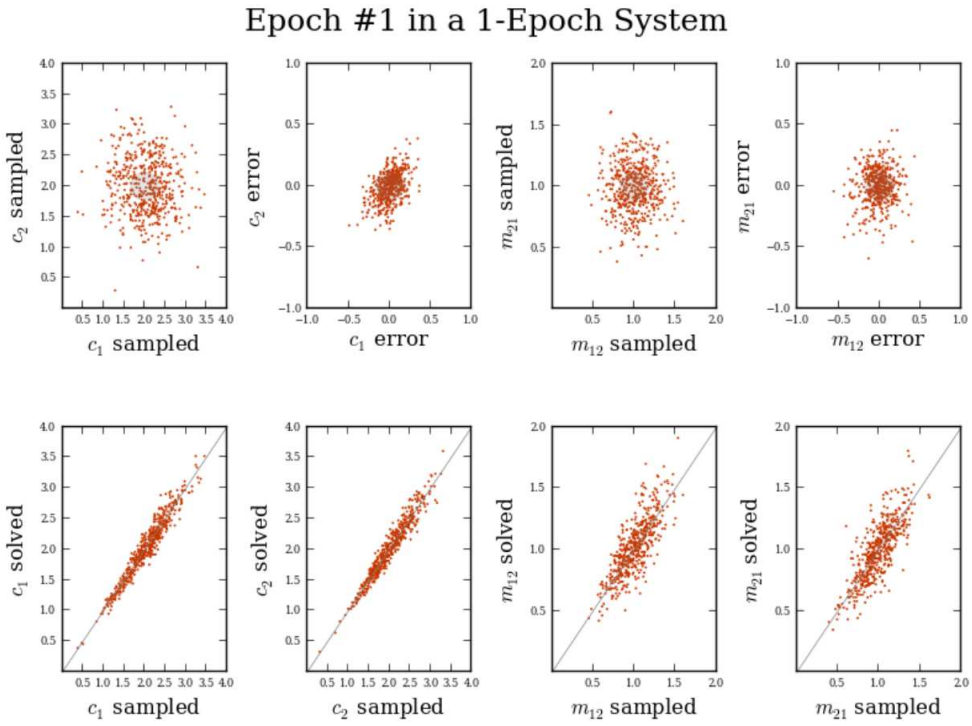

Figure 17: Population History Inference on a Single-Epoch System

The details of the eight plots are explained in Table 3. The same description applies to the rest of the experiments presented in this report.

Row #1 Plot #1		Row #1 Plot #2	
X axis	generated c_1	X axis	The difference between sampled and solved c_1
Y axis	generated c_2	Y axis	The difference between sampled and solved c_1
Expect	a circular Gaussian cloud centered on 2.0 and 2.0	Expect	a tight circular Gaussian cloud centered on 0.0 and 0.0
Row #1 Plot #3		Row #1 Plot #4	
X axis	generated m_{12}	X axis	The difference between sampled and solved m_{12}
Y axis	generated m_{21}	Y axis	The difference between sampled and solved m_{21}
Expect	a circular Gaussian cloud centered on 1.0 and 1.0	Expect	a tight circular Gaussian cloud centered on 0.0 and 0.0
Row #2 Plot #1		Row #2 Plot #2	
X axis	generated c_1	X axis	generated c_2
Y axis	solvedc_1	Y axis	solved c_2
Expect	cluster along $y = x$ with high density around $x = 2.0$	Expect	cluster along $y = x$ with high density around $x = 2.0$
Row #2 Plot #3		Row #2 Plot #4	
X axis	generated m_{12}	X axis	generated m_{12}
Y axis	solved m_{21}	Y axis	solved m_{21}
Expect	cluster along $y = x$ with high density around $x = 1.0$	Expect	cluster along $y = x$ with high density around $x = 1.0$

Table 3: Plot Details

3.3.2 Population History Inference on a Multi-Epoch System

An IM model with multiple epochs is parameterized by the coalescent rates of lineages and the rates of migration between populations for each epoch. Figure 18 is an illustration of such a model.

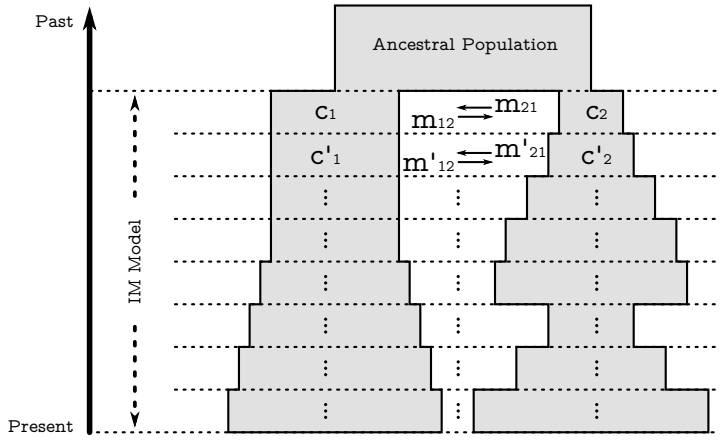

Figure 18: Isolation-with-Migration Model with Multiple Epochs

Instead of four parameters to infer, in a multi-epoch system, there are $4 \times k$ parameters to infer, where k is the number of epochs in the system. The inference procedure was tested on a three-epoch system with different initial conditions. Randomly generated values were used as the parameters to solve and are shown below. The value i denotes the epoch number backward in time.

$$\begin{aligned} c_1[i] &\sim \mathcal{N}(2.0,\, 0.5)\,, \forall i \\ c_2[i] &\sim \mathcal{N}(2.0,\, 0.5)\,, \forall i \\ m_{12}[i] &\sim \mathcal{N}(1.0,\, 0.2)\,, \forall i \\ m_{21}[i] &\sim \mathcal{N}(1.0,\, 0.2)\,, \forall i \end{aligned}$$

Figure 21 shows the inference performance of the $4 \times 3 = 12$ parameters in a three-epoch system. The following initial constants were used. The overall performance is less desirable on multi-epoch systems, and it is worse the further back in time the epoch is.

```
# of parameter combinations tested (# of red dots in each plot)    500
For each dot:  # of data points tested in a range                  50
For each dot:  # of time steps in epoch                            20
For each dot:  # of samples for each 11, 20, and 02 configuration  500
```

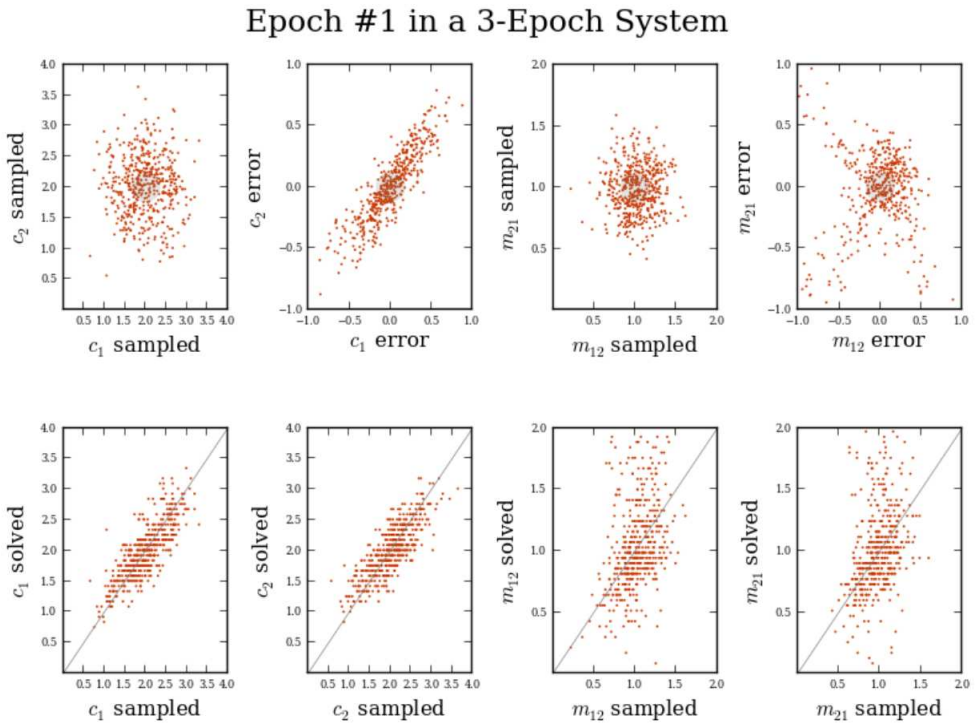

Figure 19: Inference in the Latest Epoch in a Three-Epoch System

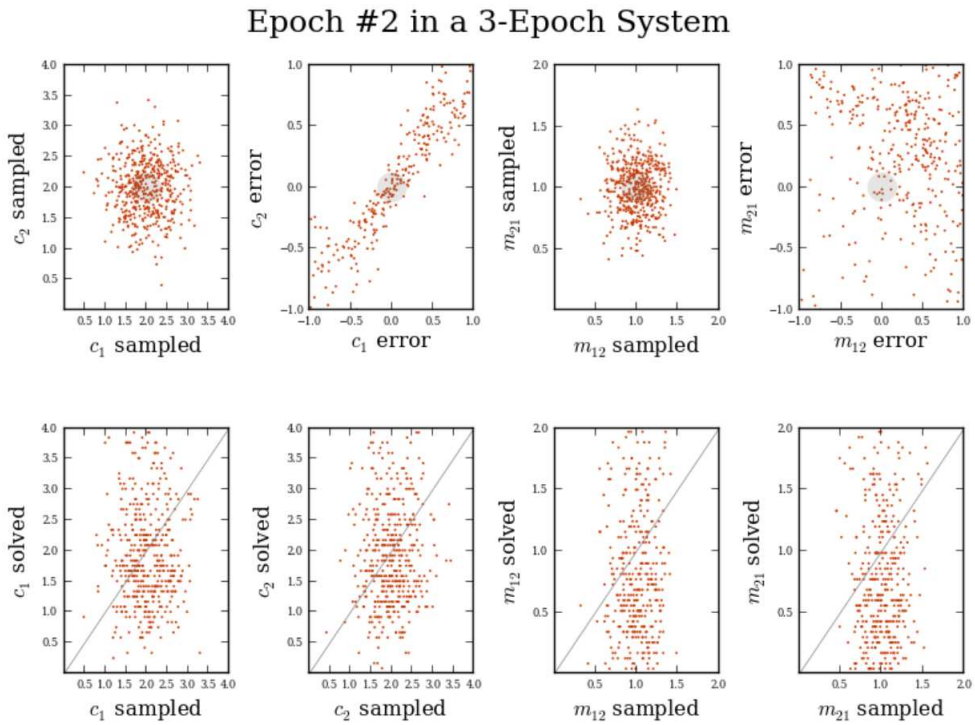

Figure 20: Inference in the Middle Epoch in a Three-Epoch System

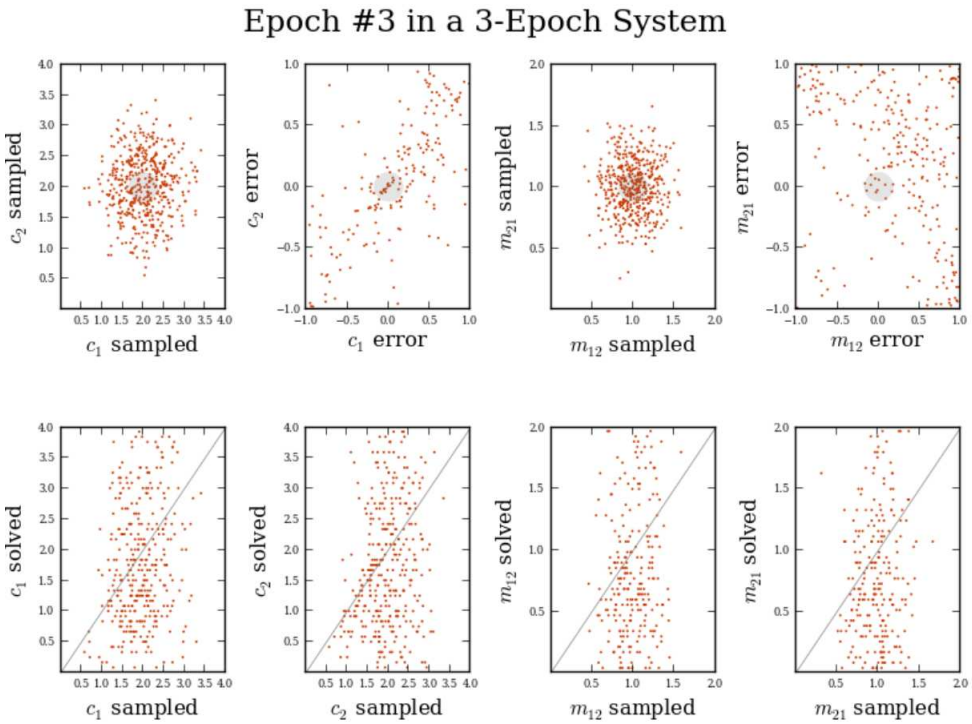

Figure 21: Inference in the Earliest Epoch in a Three-Epoch System

4 Observations and Conclusions

As we saw, the inference accuracy in a multi-epoch system is less desirable. To determine various factors that affect the inference performance, we used the one-epoch system for the following observations.

4.1 Population Size Ratio

We saw in Section 3.2.2, under the same conditions, the inference accuracy for migration rates is consistently and considerably less desirable compared to the coalescent rates. For migration events, less information is captured in the same amount of present-day genomic data since coalescent events play the dominant role most of the time. In addition, we use TMCRAs to perform the inference, and TMCRA is a direct consequence of coalescent events.

According to the nature of MLE, the most direct approach to improve this accuracy is to increase the amount of data. Section 4.2 will discuss this approach in detail. And the performance does improve as expected with a increasing volume of sample data. In this section, however, we focus on the influence of varying population parameters. Specifically, we look at the coalescent rate ratio: i.e., the population size ratio between the two isolated sub-populations. Figure 22 and 23 visualize this comparison.

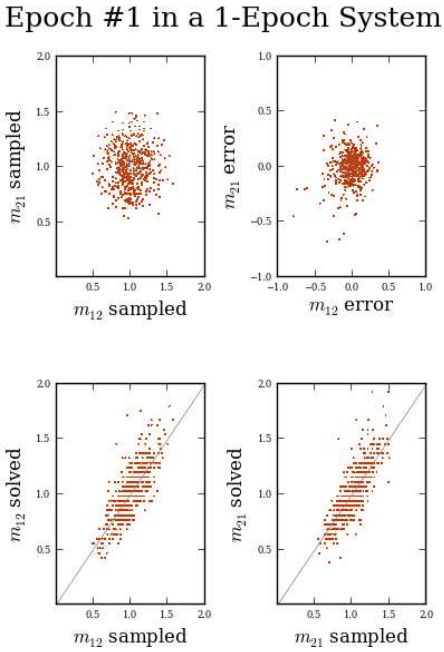

Figure 22: Population Size Ratio $1:1$

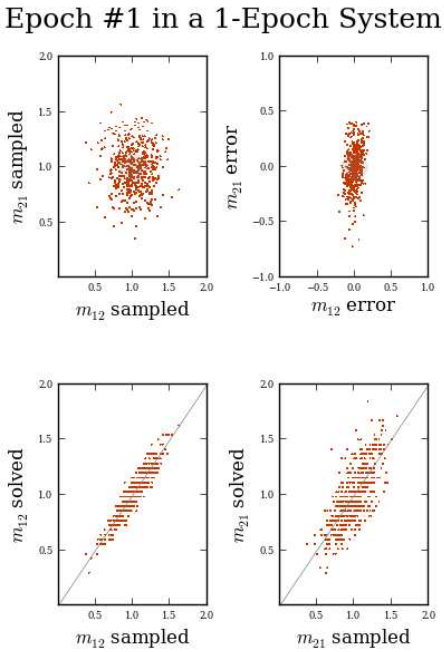

Figure 23: Population Size Ratio $10:1$

The following initial constants were used in these experiments.

```
# of parameter combinations tested (# of red dots in each plot)    500
For each dot:  # of data points tested in a range                  50
For each dot:  # of steps in each epoch                            10
For each dot:  # of samples for each 11, 20, and 02 configuration  500
```

This result is not surprising. We can interpret this phenomenon from the following two aspects.

1. Population 1 is on average 10 times the size of population 2. And migration rates in both directions are the same on average. The number of migration events from population 1 to population 2 is therefore 10 times the number of migration events from population 2 to population 1. Keeping in mind that m_{12} describes the rate of individuals migrating from population 1 to population 2, while m_{21} describes the opposite migration, within a fixed volume of sample data, on average 10 times the amount of events involving m_{12} are captured than that of m_{21}.

2. Since the population size is reversely proportional to the coalescent rate, a population that is small in size would have a fast coalescent rate. This means two individuals, e.g. the pairwise samples of a locus used in this study, find their MRCA relatively fast in population 2. With this in mind, any migration event that leads to an individual migrating from population 1 to population 2 would be more directly related to the TMRCA, which happens to be the direct measuring criteria in this study. Hence an event involving m_{12} is more informative, so m_{12} is solved with better accuracy than m_{21}.

4.2 Sample Size

The number of samples in each 1-1, 2-0, and 0-2 configuration directly influence the parameter inference. Figure 24, 25, and 26 demonstrate this situation. The following initial constants were used in these experiments. The reason the last execution has fewer data points, 150 instead of 500, is due to the extended execution time.

```
# of parameter combinations tested (# of red dots in each plot)    500,150
For each dot:  # of data points tested in a range                  50
For each dot:  # of steps in each epoch                            20
For each dot:  # of samples for each 11, 20, and 02 configuration  50,500,5000
```

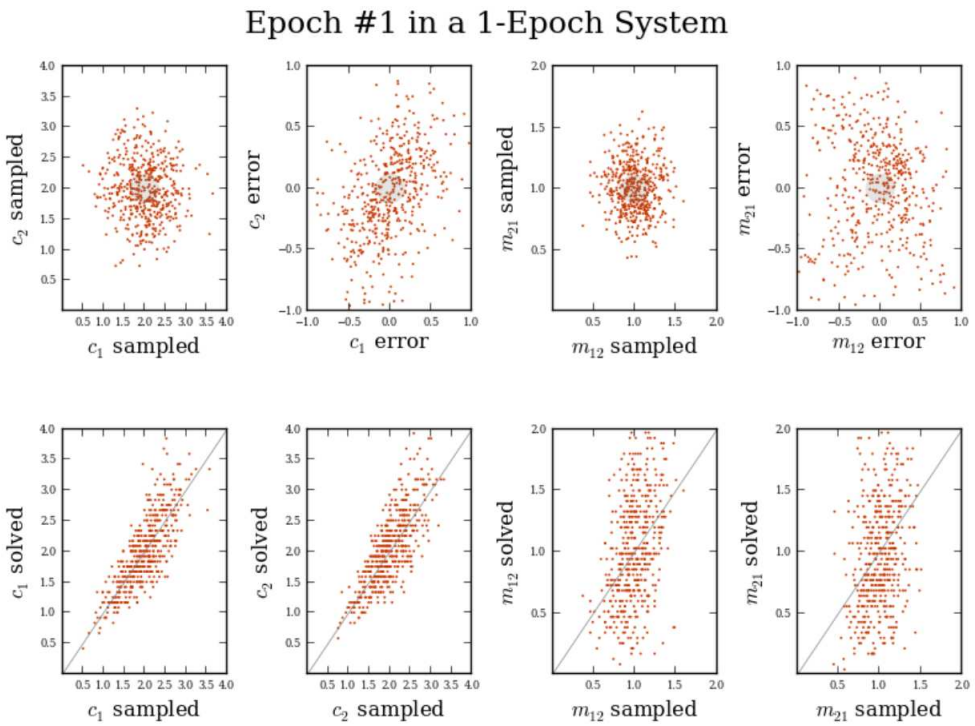

Figure 24: Sample Size of 50 for each Pairwise Configuration

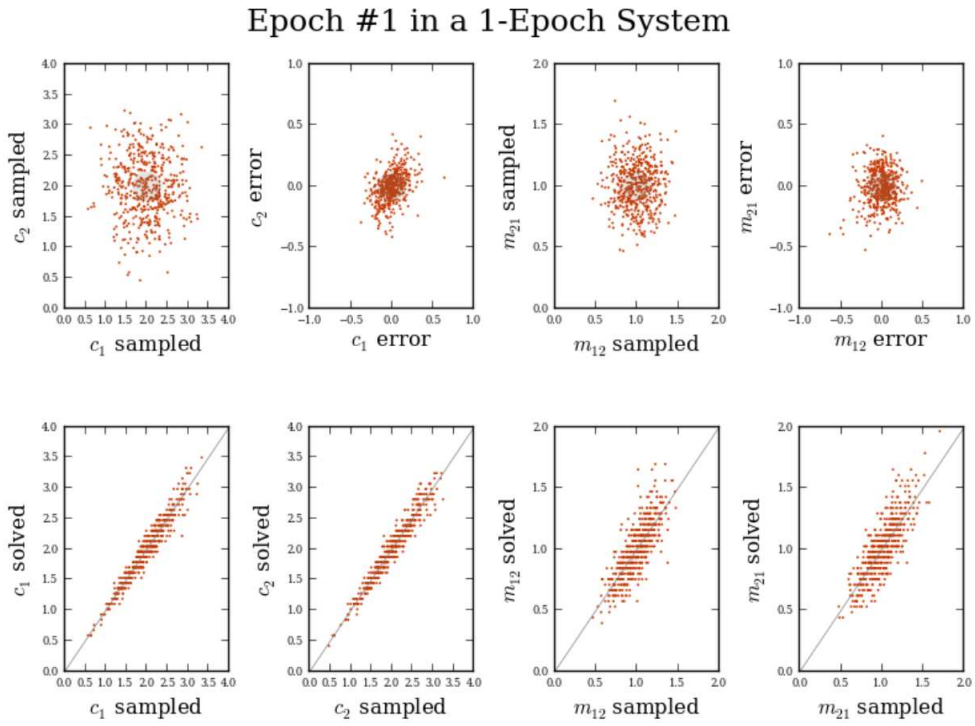

Figure 25: Sample Size of 500 for each Pairwise Configuration

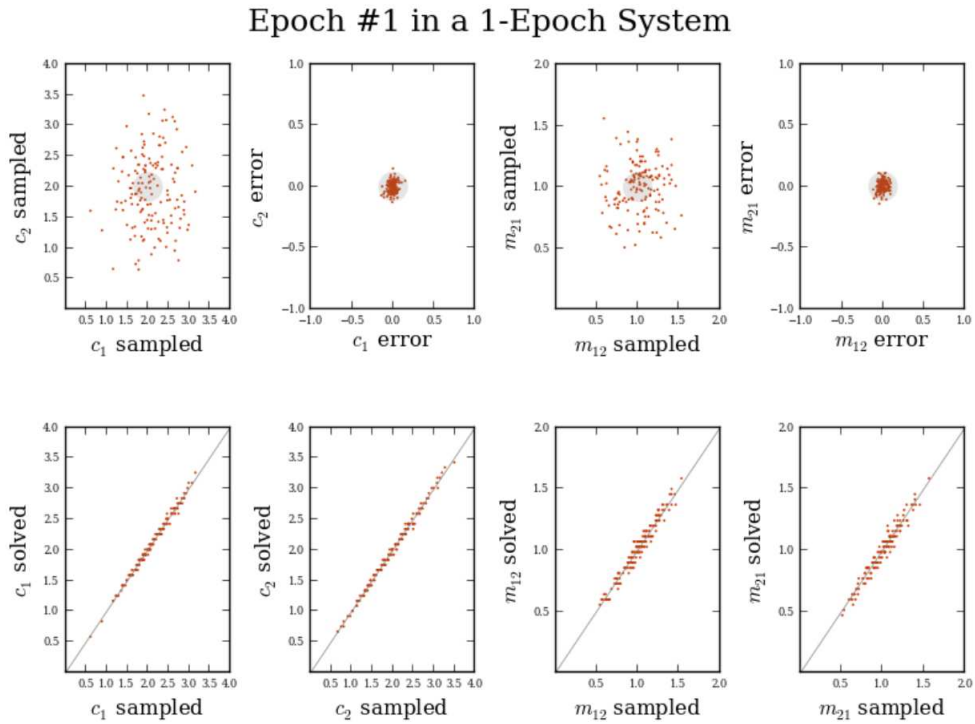

Figure 26: Sample Size of 5000 for each Pairwise Configuration

The inference accuracy increases with the sample size while the other conditions stay the same. Coalescent rates converge to their true values faster than migration rates. Refer to Figure 24 and 25. In a one-epoch system the migration rates eventually converge to their true values with a sufficient amount of data. Refer to the four plots on the right-hand side in Figure 26.

4.3 Time Steps

The discrete time steps considered in each epoch noticeably affect the inference accuracy. Figure 27, 28, and 29 demonstrate this situation. The following initial constants were used in these experiments.

```
# of parameter combinations tested (# of red dots in each plot)    500
For each dot:  # of data points tested in a range                  50
For each dot:  # of steps in each epoch                            2,6,20
For each dot:  # of samples for each 11, 20, and 02 configuration  500
```

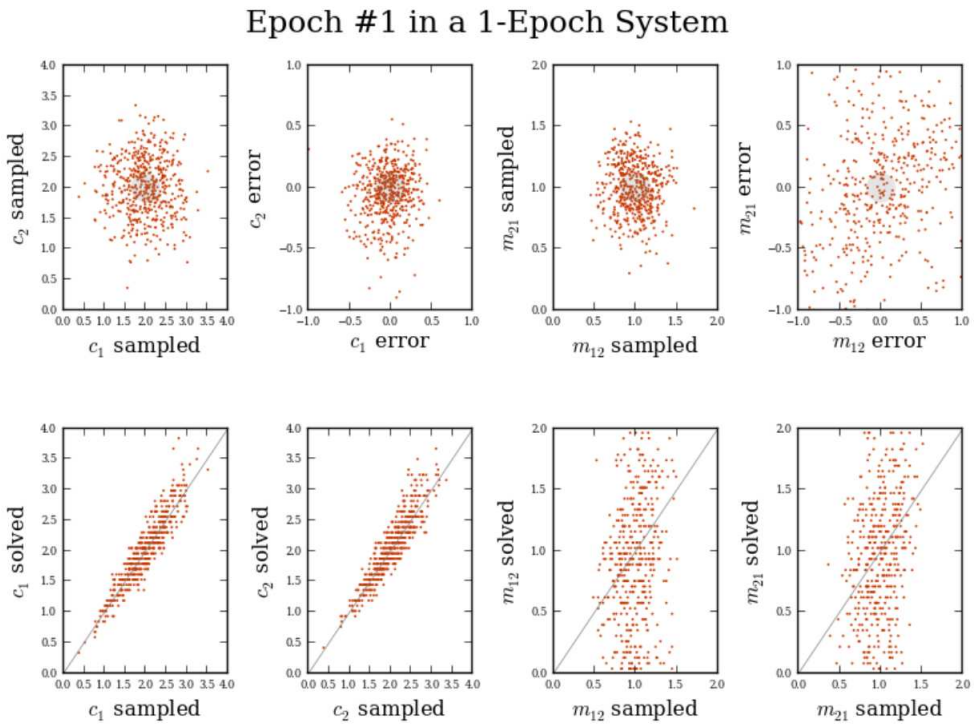

Figure 27: Time Step of 2 in Each Epoch

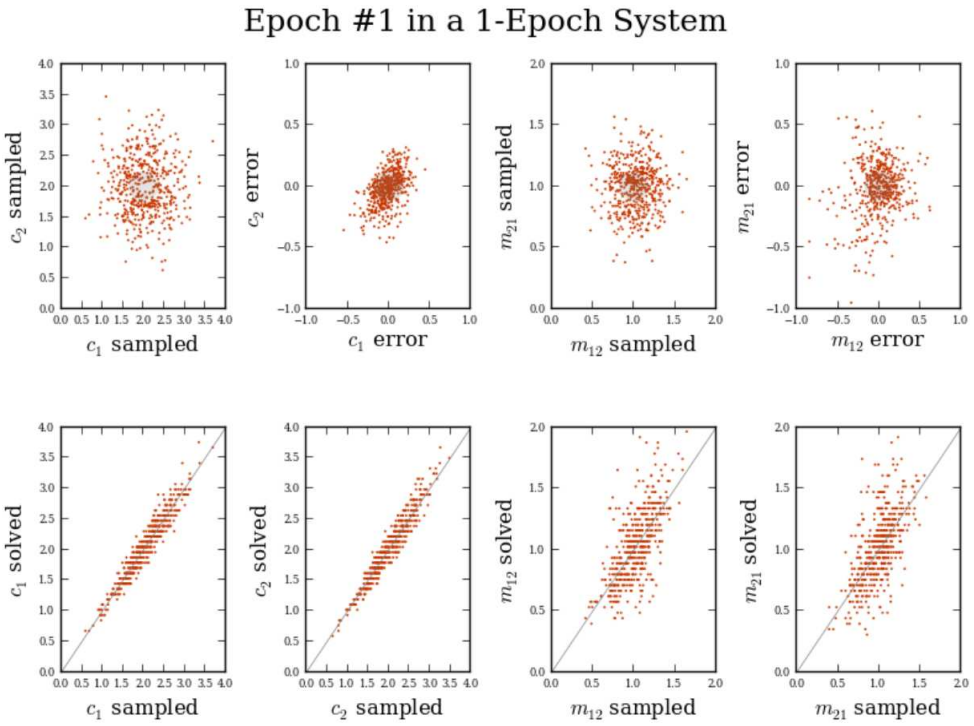

Figure 28: Time Step of 6 in Each Epoch

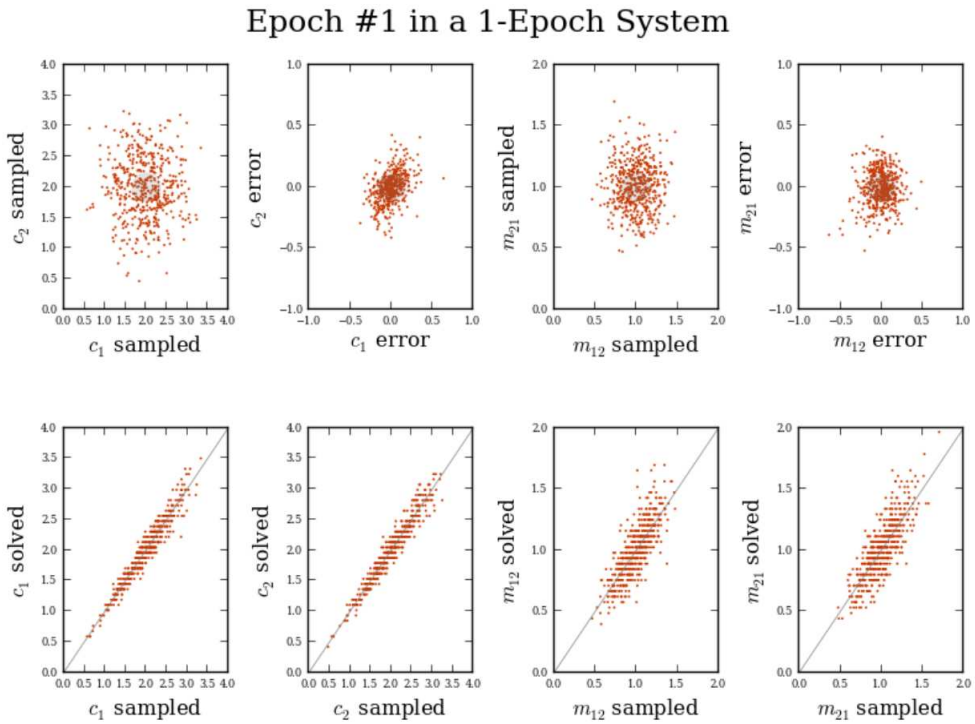

Figure 29: Time Step of 20 in Each Epoch

The inference accuracy increases with the number of discrete time steps in each epoch while the other conditions stay the same. In a one-epoch system, the migration rates converge to their true values slower than the coalescent rates; i.e., comparing plot 2 with 4 on both row 1 and 2 in Figure 27, 28 and 29.

4.4 Tested Data Points in a Range

The number of tested data points for each solved parameter does not change the shapes of the resulting plots, but when this number is significantly bigger the plots would look smoother. Figure 27 and 28 demonstrate this situation. The following initial constants were used in these experiments.

```
# of parameter combinations tested (# of red dots in each plot)    500
For each dot:  # of data points tested in a range                  50,200
For each dot:  # of steps in each epoch                            20
For each dot:  # of samples for each 11, 20, and 02 configuration  500
```

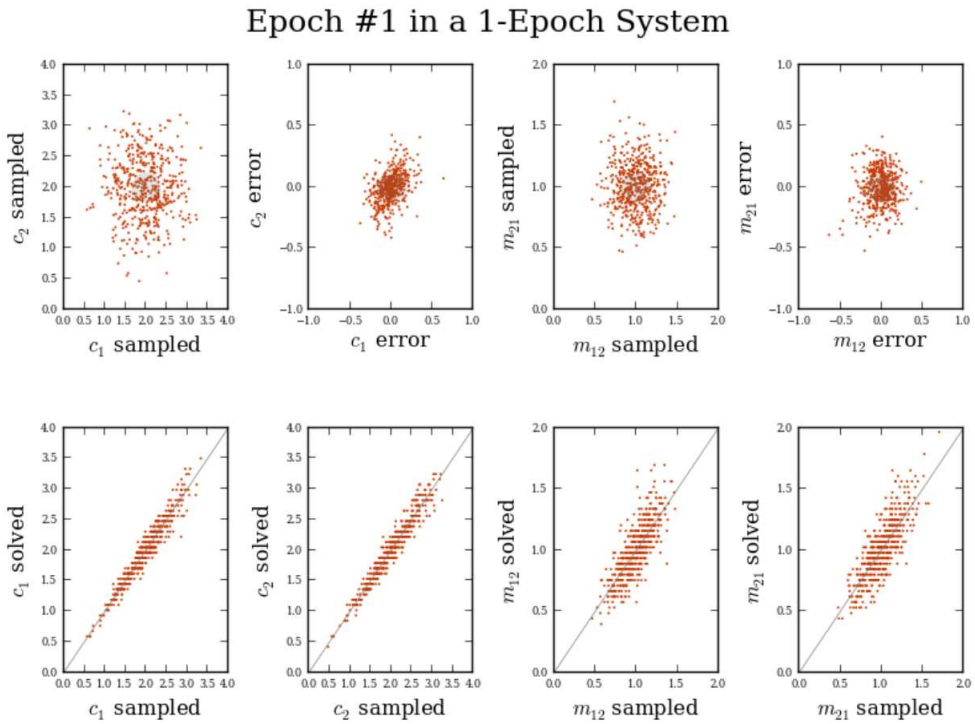

Figure 30: 50 Data Points Tested in Each inference

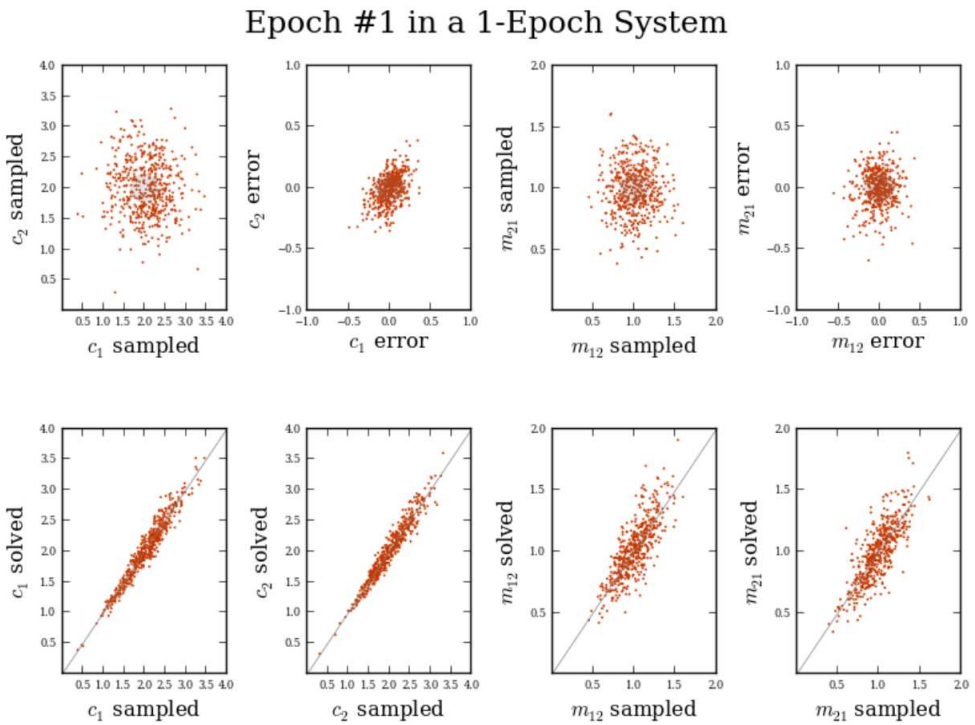

Figure 31: 200 Data Points Tested in Each inference

This condition is worth considering when the execution time becomes an issue. Arriving at this observation is beneficial for future analysis because we would most likely be interested in the overall shapes of these scatter plots rather than the accuracy of a particular data point.

5 Discussions and Future Work

In this section we briefly discuss two aspects of the future directions for this project, the main framework of the inference procedure, and improvements regarding the implementation and execution. The grand goal is to construct an effective mechanism and create a usable tool that generates accurate estimations of population parameters and hence outline the demographic history between two sub-populations in the evolution process.

5.1 Main Framework of the Inference Procedure

5.1.1 True Genealogy

In this study, we consider each coalescent event of a pairwise sample as an independent event. Each randomly sampled pairwise TMCRA does not have any knowledge of other pairwise TMCRAs, nor does it have any knowledge of the true genealogy of all lineages sampled from the two isolated populations. In reality there exists one and only one true phylogeny for the lineages sampled from these two isolated populations. Researchers are not aware of this true phylogeny, and the sample TMCRAs' indepencency is one of the assumptions from the model used in this study. With this in mind, it would be interesting to see whether or not this effect is noticeable or even play a more influential role than we expect when applying this inference method on real genomic data collected from present-day populations.

5.1.2 Multi-Epoch System

As shown in Section 3.3.2, the presented inference procedure failed to achieve a good performance on a three-epoch system, especially for epochs that are further back in time. Different sampling techniques could be considered and evaluated for making population history inference on multiple-epoch systems. Ideally, with enough epochs fully described in terms of the coalescent rates and migration rates, we can reveal a detailed demographic history of the two isolated populations.

5.1.3 Parameters Described as Functions

Let us continue with the thought of inferring population history in multi-epoch systems. The crux of the matter is to solve a set of parameters using the current-day genomic data. Due to the large number of parameters that are waiting to be solve, the runtime complexity and computational requirements of the presented inference mechanism increases dramatically. To reduce the problem into a more manageable size, we could consider modeling the parameters as functions of time. The possibility of this modification needs further investigation.

5.2 Software Application

5.2.1 Implementation

All experiments and observations described in the previous sections of this report were conducted in a serial fashion on a single processor. Time constraints were a limiting factor on the initial conditions we were able to use. For example, in Section 3.3.2 we conducted an experiment with a certain set of initial conditions on a three-epoch system. We implied that the observed inference accuracy is unacceptable. We used a sample size of 500, which achieves a median level of accuracy in a one-epoch system. Refer to Figure 25 and 26 for a performance comparison of sample size 500 versus 5000 in a one-epoch system. In a multi-epoch system, epochs share the sample data. In other words, a sample size of 500 in a three-epoch system is roughly comparable to a sample size of 150 in a one-epoch system. And in Figure 24 we have seen how poor the performance is when the sample size is insufficient. So, concluding the presented system does not solve multi-epoch systems is premature.

To infer a three-epoch system, a larger sample size is needed. In the current implementation, however, the time complexity becomes an issue for analysis with a larger sample size. The execution time for the experiment described in Section 3.3.2 lasted a few days on an Intel Core i7-2620M CPU at 2.70GHz. The implementation does not scale well with larger sample sizes, discrete time steps in each epoch, or tested data points in a range.

One simple way to reduce execution time is to parallelize the log likelihood calculations. Each set of parameters used to compute the log likelihood executes independently from the rest of the parameter sets. After a large number of parameter sets are tested, the maximum log likelihood is summarized by comparing the returned results from each execution on a parameter set. It should be relatively straightforward to turn this serial implementation into a multi-process procedure that takes advantage of multi-core CPUs or GPUs.

5.2.2 Execution

If a multi-threaded application as described in Section 5.2.1 were developed, it seems the Genome DK cluster residing in the Bioinformatic Research Centre would be an ideal platform to conduct further analysis of the presented inference mechanism, especially the inferences for multi-epoch systems.

References

[1] S. Sheenhan, K. Harris, Y.S. Song (2013) “Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach”, Genetics 112: 149096

[2] Y. Wang, J. Hey (2010) “Estimating divergence parameters with small samples from a large number of loci”, Genetics 184: 363–379

[3] R. Nielsen, J. Wakeley (2001) “Distinguishing migration from isola- tion: a Markov chain Monte Carlo approach”, Genetics 158: 885–896.

[4] J. Felsenstein (1988) “Phylogenies from molecular sequences: infer- ence and reliability”, Annu. Rev. Genet. 22: 521–565.

[5] T. Mailund, A. Halager, M. Westergaard, J. Dutheil, K. Munch, et al. (2012) “A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species”, PLoS Genetics 8: 1003125.

[6] A. Hobolth, L. N. Andersen, T. Mailund (2011) “On Computing the Coalescence Time Density in an Isolation-with-Migration Model with Few Samples”, Genetics 110: 124164

[7] R. R. Hudson, (2002) “Generating samples under a Wright-Fisher neutral model”, Bioinformatics 18:337-8

[8] C. Moler and C. Loan (2003) “Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later” SIAM Review, 45(1):3-49.