Bayesian Parameter Estimation of Hidden Markov Models

Patrik Mirzai

Master's Thesis in Statistics
Stockholm, Sweden
Spring 2019

Department of Statistics
Stockholm University

Supervisor:
Pär Stockhammar

No. 2019:06

Bayesian Parameter Estimation of Hidden Markov Models

Patrik Mirzai

Stockholm
University

Abstract

It is important in product development to be able to analyze manufacturing data so that one may draw conclusions about the performance of the system. A hidden Markov model (HMM) with a Bayesian approach is presented in order to analyze manufacturing data with regime changing properties. The method produces parameter estimates, such as the mean and the variance, within the different regimes. Bayesian HMMs also produce posterior distributions of the parameters in the different regimes. Two data sets from an electronics manufacturing system are used in this paper. Data is analyzed before the theory of Bayesian HMMs is presented. A Markov chain Monte Carlo (MCMC) simulation technique is used to estimate the parameters within the different regimes. Since the number of regimes in the data is unknown, a model comparison of several models is presented. A discussion about the assumptions made about the data and the potential problems that may arise during the parameter estimation is also included.

Keywords: Hidden Markov models, Bayesian inference, Gibbs sampler, Markov chain Monte Carlo, Semi-supervised learning, Computational statistics

"Such axioms, together with other unmotivated definitions, serve mathematicians mainly by making it difficult for the uninitiated to master their subject, thereby elevating its authority."

— Vladimir Arnold

Acknowledgments

I would like to express my deepest gratitude to Dr. Gustaf Mårtensson for the tremendous amount of support during this project. Your commitment to excellence, leadership abilities, and, most importantly, your devotion has served as an inspiration throughout this process.

I am also heartily indebted to Professor Tobias Rydén, who has guided me through the theoretical parts of this project. I am inspired by your artistry and ability to break down and solve complex problems. The completion of this thesis would not have been possible without your support. For that, I am truly grateful.

Next, I need to thank my thesis supervisor, Dr. Pär Stockhammar. Your ability to see the whole picture has helped me to structure my work. I am thankful for your guidance during this process, especially since you took a chance on this project.

One of the most difficult parts of this project has been to reconcile the engineering process with hidden Markov models. I therefore would like to thank Professor Daniel Thorburn, who through a great deal of artistry has helped me navigate through this process.

I am also very grateful for all the support I have received from Andreas Bergström, Senior Engineer at *Mycronic AB*. Your pedagogical ability to explain complex manufacturing systems can only be surpassed by your great humor and loyalty – you are a true Cobra Kai!

Thanks also to PhD candidate Parfait Munezero and Professor Mattias Villani for their invaluable comments and second opinions along the way.

This thesis would not have been written without allocated resources and supplied data from *Mycronic AB*. For that, I am grateful.

Finally, the last few years at the Department of Statistics have been the most intense years of my life so far, but also the most rewarding from a learning perspective. I am very thankful for having had the pleasure of meeting true A-players, including fellow classmates, PhD candidates and the people within the academic staff. Due to lack of space, I will not mention you all – but you know who you are!

Contents

1 Introduction

Machines that are used for the manufacturing of products sometimes exhibit a heterogeneity in the production process. For instance, items that are produced by a machine during certain periods of the manufacturing process may be similar to each other and have certain types of characteristics, i.e. the physical properties of the items are similar to each other. During other periods of the manufacturing process, the items that are produced by the same machine may also be similar to each other, but have different characteristics. The underlying process that determines the characteristics of the items is referred to as the *regime* or the *state* of the system. Some manufacturing systems may alternate between several states during the manufacturing process.

Knowing the underlying states of the system can be useful if one wants to analyze different patterns in data, as well as the characteristics of the items in each state. However, a problem that can arise in some manufacturing processes is that the underlying states that determine the characteristics of the items are not directly observable, and thus can be viewed as a hidden process.

As an example, assume that a manufacturing company deposits a special type of liquid onto a material and that the volume of the liquid is measured during the application process. Typically, there is some natural variation in the volume of the applied liquid. The volumes of the deposited fluids follow a certain statistical distribution with some mean and some variance. However, if there is more than one state present during the application process, data may be over-dispersed relative to a single distribution. For instance, if there are two states present in the system, the volume of the liquid may be a *mixture* of two distributions. Each time the liquid is applied onto the material, either state 1 or state 2 governs the distribution of the volume. Furthermore, this underlying process is considered to be hidden, i.e. not directly observable. An illustration of the process is shown in Figure 1.

Moreover, if the states are assumed to be dependent in that the probability of a system being in a certain state depends on the state attained in the previous event, data can be viewed as coming from a *dependent mixture model.* This motivates the use of *hidden Markov models*, if the states are assumed to take values in some finite set, S, called the *state space.*

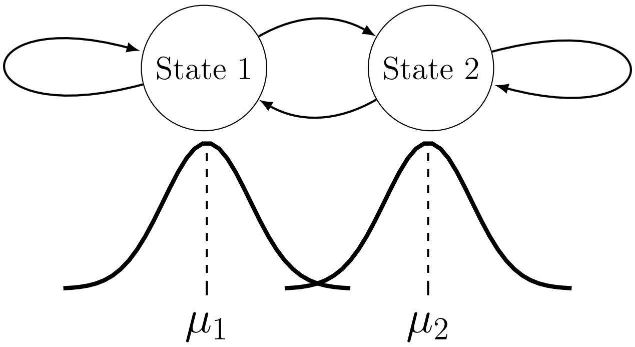


Figure 1: *An illustration of the underlying states of the system in our example. The distribution of the volume of liquid applied by the system in this example is determined by either state* 1 *or state* 2*. The nodes (circles) represent the two different states, and the edges (arrows) illustrate the process of transitioning to a state in the manufacturing process. The two distributions below the states represent the hypothetical distribution of the volume of the liquid in each state. In this case, the distribution of the volume of the liquid is Normal with the same variance in both states, but with different means.*

1.1 Background

Companies that are in the process of developing new products or systems would like to gain more information about their specific systems by analyzing manufacturing data. In this paper, we analyze a system that applies a functional material, for example, *solder paste*, onto a *printed circuit board.* Solder paste consists of a metal alloy powder and an organic resin-based *flux* that acts as an adhesive during the placement of surface mount components onto conductive *pads* on printed circuit boards. The process of transferring solder paste from the machine to a printed circuit board through non-contact deposition is called *jet printing*, or sometimes *jetting*, and the process can be seen as a modified inkjet printing technology (Mirzai and Mårtensson, 2019). Figure 2 shows a schematic of the process of jet printing of solder paste. The whole component that transfers the solder paste is called an *ejector.* In the case of solder paste, after the subsequent mounting of electronic components, the material is heated to a temperature at which the pastes metal alloy melts, wets to the component and the pad, and then solidifies, in order to connect components to the electronic *substrate.*

Each time a solder paste droplet is applied onto the printed circuit board, the volume of the droplet is measured accurately, and a common unit of measurement is nanoliters (nl). The measurements of the volume are performed by projecting a laser line pattern from an angle and measuring the distortion of the lines over the object using a 2-D camera. We assume that the measurement error is small in relation to the variation between the different droplets. The measured volume is the main characteristic used to analyze the system throughout this paper. Furthermore, the desired volume of the solder paste is referred to as the *reference.* A high-quality result of the jetting process means that the volume of the droplets are close to the reference. Although the volumes of the droplets are intended to be as close to the reference as possible, an acceptable limit for the volume commonly used in industrial applications is ±50% of the specified reference.

There are some factors that are known to affect the quality of the jetting. For instance, the manufacturing process of the ejector is very complex and each ejector has its own individual properties. These individual properties will have an effect on the characteristics of the resulting deposits. Other factors known to affect the quality of jetting are the type of solder paste used and custom hardware and software settings of the system.

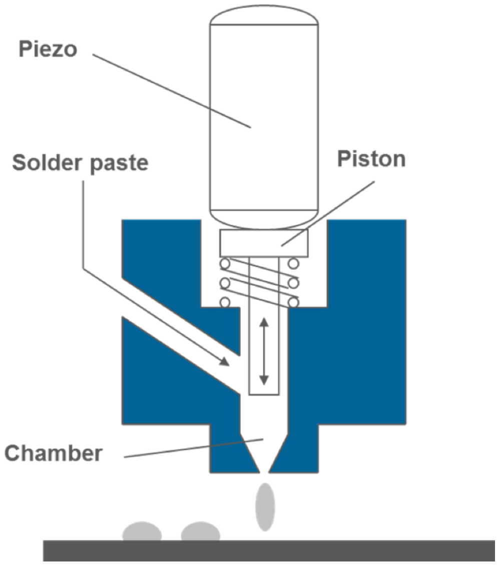


Figure 2: *A simplified illustration of the jetting process, where the ejector transfers the solder paste onto a printed circuit board.*

1.2 Problem

The main approach that is used to evaluate the system is to analyze the data from the jetting process, ranging from deposit diameter, volume, positioning, etc., although the volume is the only variable that is analyzed in this paper. Developers of the system can use the data in order to:

- Compare the performance of a new hardware or software change in the system to the initial settings of the system.
- Compare the quality of different types of solder pastes.
- Gain more information about the underlying behavior of the system.

A challenge facing these goals is to use an adequate statistical model that addresses some of these areas mentioned above. Moreover, a problem that can arise during jetting is that there can be temporary periods of irregular behavior that are reflected in the quality of the solder paste droplets. Two different data sets with different regime changing properties are therefore analyzed. The data sets consist of n measurements of the volume of solder paste deposits, $\boldsymbol{y}_{1:n} = (y_1, \ldots, y_n)$. We assume that the deposits are not independent and identically distributed (i.i.d.). Furthermore, the assumption is made that there is some periodicity in data, although it has a relatively small effect on the quality of the droplets. In addition to the previous assumptions, we also assume that there are several states or regimes that govern the distribution of the volumes. Every time a droplet is jetted onto a printed circuit board, the system is considered to be in a certain state, X_k. The distribution of the volumes of the droplets is assumed to differ between the different states. Throughout this paper, the distribution of the volume of the solder paste deposit, Y_k, conditional on the current state of the system, X_k, is assumed to be Normal with a certain mean and a certain variance. Hence,

$$Y_k | X_k = j \sim N(\mu_j, \sigma_j^2)\,, \quad j \in \{1, 2, \ldots, m\}\,, \tag{1}$$

where m denotes the total number of states.

Since the underlying states that govern the distribution of the deposits are not observable, a hidden Markov model is introduced. The dependence structure of an HMM is represented in Figure 3. The nodes represent the random variables, and the edges represent the structure of the joint probability distribution. Figure 3 implies that the conditional distribution of a state X_k, given the past states, $X_1, \ldots, X_{k-1}$, only depends on the most recent state, X_{k-1}. Similarly, the distribution of Y_k, conditional on the past observations $Y_{k-1}, \ldots, Y_1$ and the past values of the states $X_k, \ldots, X_1$, is determined by X_k only.

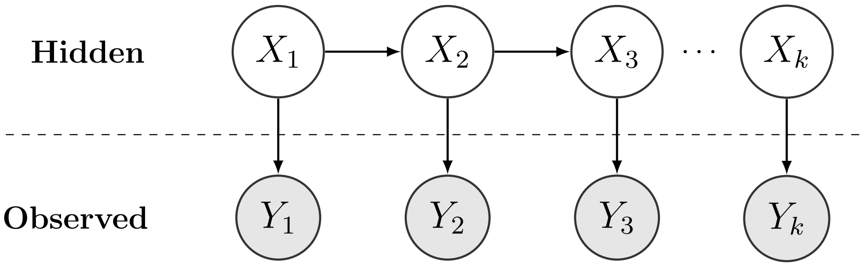


Figure 3: *A graphical representation of a hidden Markov model, where $\{X_k\}$ is the hidden chain and $\{Y_k\}$ is the observable process.*

HMMs can be used to estimate the parameters such as the mean and the variance in the different states, the probability of transitioning from one state to another (transition probability), and to gain more information about the underlying behavior of the system. Moreover, the number of hidden states in the system has to be specified when estimating the parameters in the HMM. Hence, some problems considered in this paper include; an estimation of the parameters in the different states, an estimation of the transition probabilities, and a comparison of several probable models.

1.3 Theoretical Background

Mixture models are commonly used to estimate the parameters in the different states when the underlying states form an i.i.d. sequence. See e.g. Feller (1943) or Frühwirth-Schnatter (2006) for literature on this topic.

Moreover, HMMs are commonly used when the underlying process that determines the distribution of the volumes is not an i.i.d. sequence. Some application areas of HMMs include speech recognition (Rabiner, 1989) and handwriting recognition (Makhoul et al., 1994). The theory of HMMs goes back to at least Baum and Petrie (1966), as well as a seminal version of the *EM algorithm* for HMMs, introduced by Baum et al. (1970). The algorithm can be described as a method to compute a point estimate. More recent literature on HMMs include Cappé et al. (2005), Rydén (2008) and Zucchini and MacDonald (2009), as well as seminal references therein.

1.4 Purpose

The aims of this paper are the following:

- To analyze if HMMs can serve as a tool to describe regime changes during the jetting process.
- To estimate the parameters in each state using Bayesian HMMs.
- To analyze and discuss possible problems that may arise during the parameter estimation of HMMs.

Note that we are not interested in drawing any major conclusions about the hardware or software settings of the system based on the data that is used in this paper, but rather in analyzing if HMMs can be used within jet printing, and the possible problems that can arise when using the method.

1.5 Contributions

A common method to analyze data that stems from jet printing is to use *descriptive statistics*, such as the mean and the variance together with box plots when measuring the thickness, or height, of the solder paste deposits (Huang et al., 2011; Kay et al., 2014). Moreover, a *frequentist* approach using *hypothesis testing* has been carried out by Mirzai and Mårtensson (2019) in order to determine if there is a significant difference between the mean of the volumes of the solder paste droplets when introducing a new feature to the ejector, in this case, two variants of the printing mechanisms. The study shows that hypothesis testing is unreliable in the case of jet printing. To the best of our knowledge, a Bayesian framework together with HMMs has not been applied within jet printing. Hence, the main contribution of this paper is to reconcile Bayesian HMMs with jet printing.

1.6 Outline

We begin by analyzing the distribution of the solder paste deposits. Then, theory connected to HMMs is presented, followed by a parameter estimation of the different states. The parameter estimation is done using an MCMC technique, which also provides a measure of the uncertainty of the estimates. The results are followed by a discussion about the potential problems that arise during parameter estimation.

1.7 Delimitations

Throughout this paper, we assume that there is a finite number of discrete states in the system. An alternative method is to use *state space models* when dealing with a continuous state space. An HMM is in fact a state space model with a finite state space. A thorough explanation of state space models can be found in Petris et al. (2009) and references therein.

Moreover, the problem of estimating the number of hidden states is discussed in this paper, although we are mainly concerned with parameter estimation of the hidden states. There are other *nonparametric* methods that avoids restrictive assumptions of parametric models, such as the *Dirichlet process*. Some literature on this topic can be found in e.g. Sudderth (2006) or Teh et al. (2006).

A final comment should be made regarding the rationale behind the choice of using HMMs. We assume that the underlying process that governs the distribution of the volume of the droplets is not an i.i.d. sequence. Furthermore, we make the assumption that the volumes of the droplets, $\{Y_k\}$, are conditionally independent given the states, $\{X_k\}$. This may be a strong assumption, and a further discussion on this topic presented in Section 6.2. *Markov-switching models* can be used if one believes that the distribution of data is not only governed by the current state of the system, but also by previous observations. See e.g. Hamilton (1989) for a famous example of a Markov switching process, called the switching autoregressive process, that was introduced to model econometric data. Moreover, if data is dependent, one could use *time series analysis* to describe the volumes of the solder paste droplets. However, due to the regime changing properties of the data, we find that HMMs are more suitable in this case.

2 Data

2.1 Description of Data

Let us now describe the jetting process and how the data is collected in more detail. Although some descriptions of the jetting process may seem rather technical, they are necessary in order to make the proper assumptions about the data.

The solder paste deposits are for *research and development* purposes jetted on a sheet of A4 photographic paper. A small sample of the jetted solder paste deposits is shown in Figure 4a. Each dot represents a solder paste droplet that has been jetted through the ejector. The red lines in Figure 4b show the paths that the solder paste deposits are jetted along, where each line is referred to as a *strip*. The order in which the strips are jetted depends on the software settings of the system.

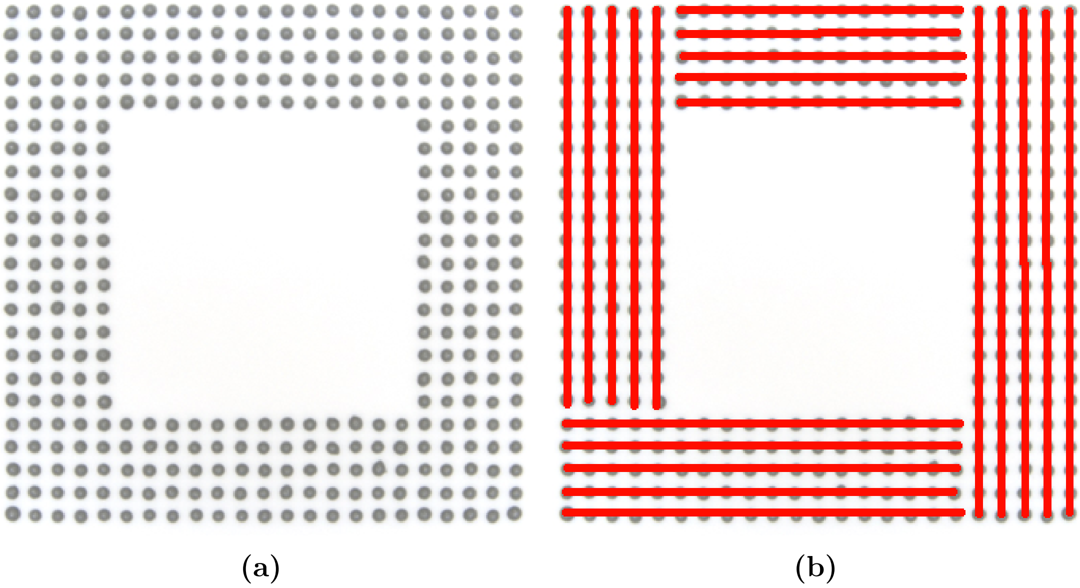


Figure 4: *Illustration of the jetting in the testing phase, where a picture has been taken from above.*

Furthermore, the dots form a rectangle, where each rectangle consists of 360 droplets. The rectangles are referred to as a generic *ball grid array* (BGA), which is a typical component application for the jet printer. The BGAs are jetted sequentially until the desired number of observations are met.

In order to ensure a stable jetting process over time, the machine has automated settings that manipulate the inputs of the system, in order to obtain the desired effect on the output of the system. In practice, this means that the initial droplets in each strip sometimes deviate from the rest of the droplets, in terms of volume. The number of initial droplets in a new strip that are affected by the settings of the machine typically range between one and three. This phenomenon is referred to as *periodicity* in data.

A plot of the volume of 40 droplets from a jetted BGA test is shown in Figure 5, where the droplets are jetted according to the reference without any disturbances in the system. The data is a subset of sequentially jetted BGAs that consist of 99 360 deposits. The strips shown in Figure 5 are of length 23. Furthermore, the reference volume is 1.85 nl, where the initial observation represents the volume of the first deposit in a strip. From the plot, it is clear that the volume of the initial droplet is smaller than the volume of the rest of the droplets in the strip. For new strips that are jetted during the jetting process, this pattern sometimes repeats itself. Observation number 24 shows a similar behavior, since it represents the first droplet in a new strip. There are also occasional deviations in the volume of the solder paste droplets, that are not caused by the settings of the machine. For instance, observation number 35 deviates from the previously jetted droplets, similar to the behavior of the first droplet in a new strip.

A further comment has to be made regarding the deviation of the initial droplet in a new strip. It is not always the case that the deviations are smaller than the rest of the droplets, as in the example shown in Figure 5. The volume of the initial droplets in a new strip can also be larger than the volumes of the rest of the droplets in the strip, and this behavior typically varies during the jetting process.

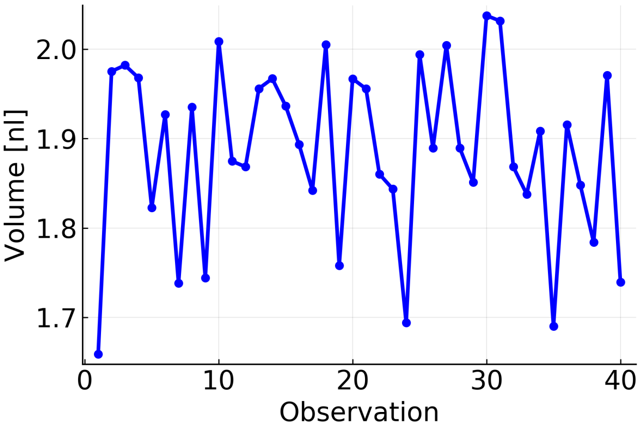


Figure 5: *A plot of the volume of jetted deposits when the droplets are jetted according to the reference. The plot shows the impact of the settings of the machine.*

It is not only the settings of the machine that affect the quality of the jetted solder paste deposits. There are many variables that can affect the quality of the droplets during the jetting process, which can result in droplets that deviate from the reference, some of which were mentioned in Section 1.1. The magnitude of the deviation varies but is often similar to the periodicity shown in Figure 5.

A recurring problem during jetting is individual or groups of solder alloy particles that can get stuck in the ejector. A common belief is that the distribution of the volume of the droplets differs during the period that particles are stuck in the ejector. The duration, in terms of the number of jetted solder paste droplets, for which particles are stuck in the system varies. It can range between one droplet to several million droplets, and during this process, the ejector can break. Different types of solder pastes are manufactured in different ways, and the risk of particles getting stuck in the system therefore varies, depending on the type of solder paste that is used.

In this paper, we can roughly categorize the data from the jetting process into two groups. The first group of data typically arise following a replacement of the ejector, due to component failure. During the initial stages of jetting following the replacement, there is a possible *initiation period* for the ejector. During this period, the solder paste deposits are typically not jetted according to the reference. Furthermore, the variance of the volume is usually large during the initiation period. The second group of data arises after the initiation period, when the system typically jets the solder paste droplets close to the reference, although there are some disturbances during the jetting. The process when the system jets in the neighborhood of the reference, with some regular disturbances, is referred to as the *system equilibrium*. We use two different data sets to analyze the different groups of data in this paper. The first data set is used to analyze the initiation period and the second data set is used to analyze the system equilibrium.

2.1.1 Initiation Period

Let us now examine a typical behavior that occurs during the initiation period of jetting, following a replacement of the ejector. A data set that is truncated from above, with reference volume 1.85 nl is used. A subset of the data set was used to illustrate the initial droplet effect in a new strip in Figure 5. The data consists of 99 360 observations of the volume of solder paste deposits. Hence, a set of 276 BGAs have been jetted sequentially. This data is shown in Figure 6.

We assume that two different states govern the distribution of the volume, where the initial 60 000 droplets, roughly, can be seen as belonging to the initiation period. The mean volume of the solder paste droplets during the initiation period does not seem to follow the reference. Furthermore, the variance of the volume seems to be very high during the initiation period, compared to the variance of the volume during the system equilibrium.

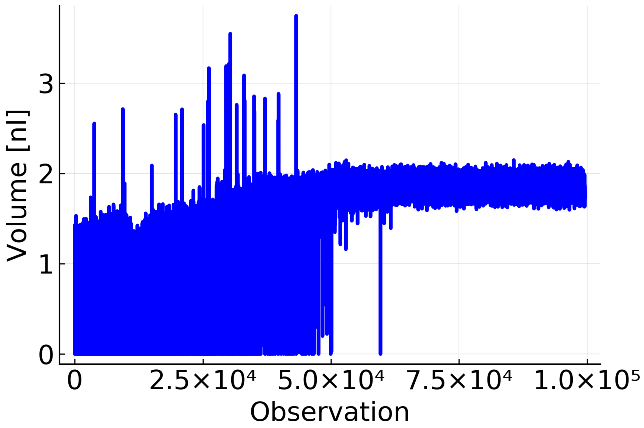


Figure 6: *Illustration of a jetting sequence including the initiation period. The initial 60 000 jetted droplets, approximately, belong to the initiation regime. After the initiation period, the distribution of the volumes is concentrated around the reference, 1.85 nl, and can be considered to belong to the system equilibrium regime.*

In this data set, we are not interested in estimating the parameters in the two different regimes. Instead, we are interested in estimating the number of droplets that are jetted before reaching the system equilibrium, which can be denoted by ξ. Estimating this quantity each time an ejector is replaced in the system can give a measure over the number of droplets that manufacturers need to jet before reaching system equilibrium, following a replacement of the ejector.

It is not unusual that there are several states present in the system after the initiation period. In that case, the mechanism is similar for obtaining the estimate of ξ. For instance, assume that there are three states present in the system. State 1 represents the initiation period, and state 2 and state 3 govern the distribution of the volumes during the system equilibrium, where the system alternates between the two states during that period. The estimate of ξ can be obtained by recording the last time point that the system is in state 1 during the jetting process.

2.1.2 System Equilibrium

After the initiation period, it is of interest to learn more about the distribution of the volume of the solder paste droplets. If there is more than one state present in the system, i.e. there is an underlying process that governs the distribution of the volumes, one is interested in estimating the parameters of this process, such as the mean and the variance of the volume in each state. The information can be used to compare different types of solder pastes. Furthermore, information about the mean and the variance in each state can be used to measure the performance of the system when introducing changes in hardware or software.

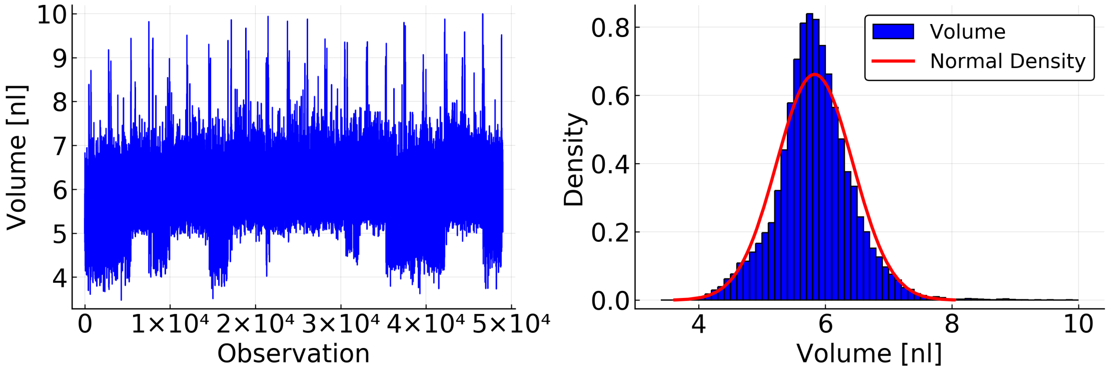
As an example, assume that the developer of the system would like to analyze the performance of the jetting process when making a hardware change and compare the performance of the system to the initial settings in the present system. A system that has states with a mean close to the reference, together with a small variance is preferred.

Another type of data set that is truncated from below and above, which consists of 49 000 observations is used. The droplets are considered to be measured during the system equilibrium, i.e. when the droplets are jetted according to the reference, in this case 5.8 nl. In this data set, the droplets are no longer jetted according to a BGA. Instead, they are organized in strips of length 280 deposits. Thus, the data consist of 175 strips. Having longer strips means that the data is less affected by periodicity, i.e. the effect from the initial droplets in a new strip. Although the initial droplet in a new strip may deviate from the rest of the droplets in the strip, the assumption is made that the probability of transitioning from a state to another does not change at that time point.

A plot of the volume of the jetted solder paste droplets in this system, in nanoliters, is shown in Figure 7a. The data appears to come from more than one state, due to the sudden regime changes in the data. Furthermore, there appears to be a pattern where the volume temporarily increases with a certain regularity during the jetting. These events are treated as random in this paper, although it can be argued that they are a part of a cyclical pattern. A discussion about this phenomenon is given in Section 6.1.

The histogram of the data described in Figure 7a is given in Figure 7b and has a similar shape as the Normal distribution, even if it does not completely follow the density function of a Normal distribution with indicated mean and variance.

Due to the nature of the system previously described, we believe that data is a *mixture* of at least two Normal distributions, during the system equilibrium. Hence we believe that there are at least two states or regimes that govern the distribution of the volume during the system equilibrium.



(a) **(b)**

Figure 7: *Illustration of the volume of the jetted solder paste droplets in nanoliters (nl), where (a) is a plot of the volumes of the droplets during one test and (b) shows the distribution of the volume for the same test. The reference volume is 5.8 nl and the red line in the histogram represents the density of a Normal distribution with the same mean and the same variance as the data set.*

2.2 Test of Normality

Let us now analyze the normality of the data during the system equilibrium, in order to assess if the volume of the droplets are generated from more than one state. There are several methods that can be used to evaluate the normality of the data, both graphically and analytically. We use a quantile-quantile (Q-Q) plot for a graphical inspection of the normality of data (Wilk and Gnanadesikan, 1968), and the Kolmogorov-Smirnov (K-S) test for evaluating the normality analytically (Feller, 1948). The same data set shown in Figure 7a and Figure 7b, with 49 000 observations, is used for the *goodness of fit* test.

A comment has to be made regarding the assumptions of the K-S test. One of the assumptions made about the data is that the observations are i.i.d., which does not hold in our case. The test is still performed, although the same interpretation as in the i.i.d. case cannot be made.

2.2.1 QQ-Plot

Assume that we have a sample of n observations, $\boldsymbol{y}_{1:n} = (y_1, \ldots, y_n)$. The QQ-plot can be obtained by plotting the theoretical quantiles, $F^{-1}(F_n(y_i))$, against the sample quantiles, $y_{(i)}$. The plot is then analyzed graphically. The points in the Q-Q plot fall on the line of identity if the empirical distribution, F_n, is consistent with the theoretical distribution function, F, (Loy et al., 2016). Plotting the theoretical quantiles against the sample quantiles, we see that the blue dots are not consistent with the line identity, i.e. the red line.

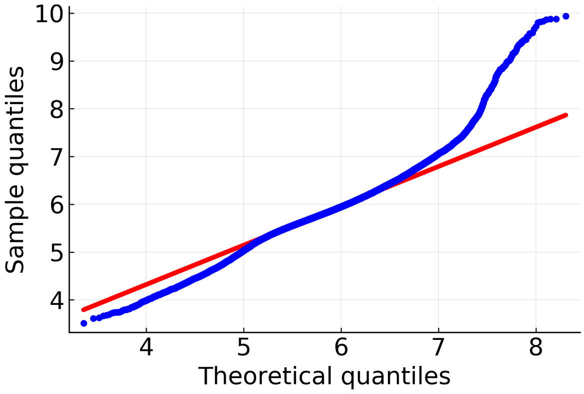


Figure 8: *A QQ plot showing the plotted theoretical quantiles against the sample quantiles, represented by the dots in blue. The red line represents the line identity.*

2.2.2 Kolmogorov-Smirnov Test

The K-S test can be used to compare the empirical distribution function, F_n, with the hypothesized distribution function, F. The test statistic, D_n, is the largest vertical distance between F_n and F. Thus,

$$D_n = \max_y |F_n(y) - F(y)| ,$$

where $\max_y$ denotes the maximum of the set of distances. Formally, the hypotheses are given by

$$\begin{aligned} H_0 &: F_n(y) = F(y) , \quad \forall\, y \in \mathbb{R}^+ , \\ H_1 &: F_n(y) \neq F(y) , \quad \text{for some } y \in \mathbb{R}^+ , \end{aligned}$$

where $\mathbb{R}^+$ denotes the set of the non-negative real numbers.

We use the 1% significance level when performing the K-S test. The critical value is approximately given by $D_{\text{crit},0.01} = 1.63/\sqrt{n}$. If $D_n > D_{\text{crit},0.01}$, the null hypothesis is rejected.

Based on the data set, our statistic is $D_n = 0.0419$, and hence, the null hypothesis is rejected at 1% level of significance, since D_n exceeds the critical value of $D_{\text{crit},0.01} = 1.63/\sqrt{49000} \approx 0.0074$.

In summary, both the QQ-plot and the K-S test imply that the data from system equilibrium does not follow a single Normal distribution, although the i.i.d. assumption is relaxed in the K-S test.

3 Hidden Markov Models

3.1 Markov Chains

In this subsection, a brief introduction to Markov chains is given. A thorough explanation of Markov chains can be found in e.g. Grimmett and Stirzaker (2001) or Ross (2010).

Let $\{X_k, k \in \mathbb{Z}^+\}$ be a *stochastic process* that takes on a finite or countable number of possible values in the state space, S. Here, $\mathbb{Z}^+$ denotes the set of all positive integers. We assume that S takes on m possible values, $m \in \mathbb{Z}^+$. The process $\{X_k\}_{k \geq 1}$ is said to be a Markov chain if it satisfies the *Markov property*:

$$P(X_{k+1} = j | X_k = i, X_{k-1} = i_{k-1}, \dots, X_1 = i_1) = P(X_{k+1} = j | X_k = i)\,, \qquad (2)$$

for all states $i_1, i_2, \dots, i, j \in S$, and for all $k \geq 1$. Furthermore, a Markov chain is said to be *homogeneous* if:

$$P(X_{k+1} = j | X_k = i) = P(X_2 = j | X_1 = i) = \gamma_{ij}\,, \quad k \in \mathbb{Z}^+\,,\ i, j \in S\,,$$

where γ_{ij} denotes the *one-step transition probability*. Note that $\gamma_{ij} \geq 0\ \forall\, \gamma_{ij}$ and $\Sigma_{j=1}^{m} \gamma_{ij} = 1$. We assume that the Markov chains we are dealing with in this paper are homogeneous throughout, unless otherwise stated.

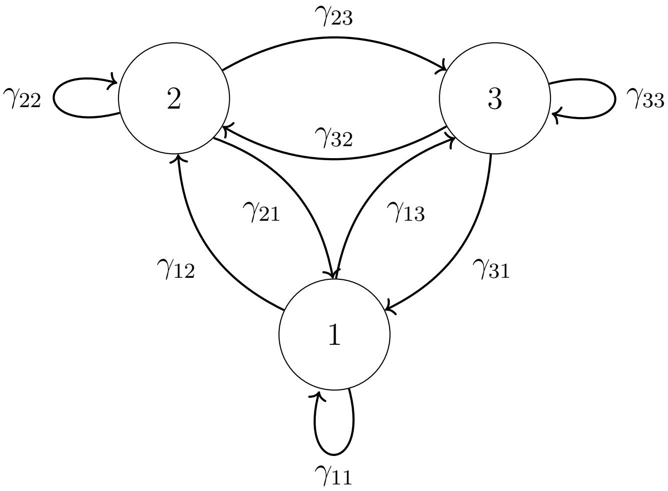


Figure 9: *A Markov chain with three states. The nodes represent the different states, and the edges together with the γ's represent the one-step transition probabilities.*

It is often convenient to express the one-step transition probabilities γ_{ij} in a matrix. Thus, $\mathbf{\Gamma}$ is defined as the matrix with (i, j) element γ_{ij}:

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1m} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m1} & \gamma_{m2} & \cdots & \gamma_{mm} \end{pmatrix}.$$

Here, $\mathbf{\Gamma}$ is a matrix of size $(m \times m)$, where m denotes the number of states in the Markov chain.

Moreover, the n-step transition probabilities for a homogeneous Markov chain are defined as

$$\gamma_{ij}(n) = P(X_{k+n} = j | X_k = i) \,,$$

and expressed in a matrix, $\mathbf{\Gamma}(n)$, similar to the one-step transition probability matrix.

An important property of all finite state-space homogeneous Markov chains is that they satisfy the *Chapman–Kolmogorov equations*, which in matrix notation can be written as:

$$\mathbf{\Gamma}(n+k) = \mathbf{\Gamma}(n)\mathbf{\Gamma}(k) \,.$$

The proof is omitted here, but can be found in e.g. Section 6.1 in Grimmett and Stirzaker (2001). The Chapman-Kolmogorov equations imply that

$$\mathbf{\Gamma}(k) = \mathbf{\Gamma}(1)^k, \quad \forall\, k \in \mathbb{Z}^+ \,.$$

Moreover, the *unconditional* probabilities of a Markov chain, $P(X_k = j)$, are often of interest. That is, the probability of being in state j at a given time k. These probabilities can be expressed as a row vector of length m:

$$\boldsymbol{\pi}(k) = (P(X_k = 1), \ldots, P(X_k = m)), \quad k \in \mathbb{Z}^+ \,.$$

To compute the unconditional probabilities at time $k+1$, we simply postmultiply the unconditional probabilities with the transition probability matrix $\mathbf{\Gamma}$:

$$\boldsymbol{\pi}(k+1) = \boldsymbol{\pi}(k)\mathbf{\Gamma} \,. \tag{3}$$

Furthermore, the *initial state probabilities*, or *initial distribution* of the Markov chain is denoted $\boldsymbol{\pi}(1)$, or abbreviated $\boldsymbol{\pi}$:

$$\boldsymbol{\pi} = (P(X_1 = 1), \ldots, P(X_1 = m)) = (\pi_1, \pi_2, \ldots, \pi_m) \,, \tag{4}$$

where $\boldsymbol{\pi}$ is a row vector of length m.

3.2 Basic Properties of Hidden Markov Models

Although a brief introduction to hidden Markov models was presented in Section 1.2, let us formally introduce its basic properties. A hidden Markov model consists of two parts. The first part is an unobservable Markov chain $\{X_k\}_{k\geq 1}$, where k is an integer index. The second part consists of an observable *stochastic process* $\{Y_k\}_{k\geq 1}$ that is linked to the Markov chain in that X_k governs the distribution of Y_k. Furthermore, the assumption is made that $\{X_k\}$ must be the only variable that affects the distribution of $\{Y_k\}$. That is, the conditional distribution of $\{Y_k\}$ given $\{X_k\}$ is a sequence of independent random variables.

The structure of the HMM can be expressed as

$$P(X_{k+1}|X_k, \dots, X_1) = P(X_{k+1}|X_k)\,, \quad k \in \mathbb{Z}^+\,,$$

and

$$P(Y_k|Y_{k-1}, \dots, Y_1, X_k, \dots, X_1) = P(Y_k|X_k)\,, \quad k \in \mathbb{Z}^+\,.$$

An alternative approach is to view the hidden Markov model as a dependent mixture model. Here, $Y_k, \dots, Y_1$ and $X_k, \dots, X_1$ represent histories from times 1 to k. Although $\{Y_k\}$ is conditionally independent given $\{X_k\}$, the observed variables $\{Y_k\}$ are themselves not an independent sequence because of the dependence in $\{X_k\}$. A thorough discussion about the structure of the HMMs can be found in e.g. Cappé et al. (2005), where the authors also discuss how the process $\{Y_k\}$ in general does not have the loss of memory property of Markov chains, i.e. that the conditional distribution of Y_k given $Y_1, ..., Y_{k-1}$ generally also depends on all the conditioning variables.

3.3 Emission Probabilities

The *emission probabilities* are defined as

$$f_j(y_k) = P(y_k|X_k = j)\,, \quad j \in \{1, 2, \dots m\}\,,\ k \in \{1, 2, \dots, n\}\,.$$

Since we assume the jetted solder paste deposits follow a Normal distribution in each state, the emission probabilities can be written as

$$f_j(y_k) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{(y_k - \mu_j)^2}{2\sigma_j^2}\right\}\,, \quad j \in \{1, 2, \dots m\}\,,\ k \in \{1, 2, \dots, n\}\,.$$

3.4 Marginal Distribution

Recall that the unconditional state probabilities are defined as $\pi_j(k) = P(X_k = j)$ for $j \in \{1, 2, \dots, m\}$ and $k \in \{1, 2, \dots, n\}$. Furthermore, the unconditional probability, $P(y_k)$, can be computed as

$$\begin{aligned} P(y_k) &= \sum_{j=1}^{m} P(X_k = j) P(y_k|X_k = j) \\ &= \sum_{j=1}^{m} \pi_j(k) f_j(y_k)\,. \end{aligned}$$

In matrix notation, the expression can be rewritten as

$$\begin{aligned} P(y_k) &= (\pi_1(k), \ldots, \pi_m(k)) \begin{pmatrix} f_1(y_k) & & 0 \\ & \ddots & \\ 0 & & f_m(y_k) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= \boldsymbol{\pi}(k)\boldsymbol{P}(y_k)\mathbf{1'}, \end{aligned}$$

where $\boldsymbol{P}(y_k)$ is defined as the diagonal matrix with jth diagonal element $f_j(y_k)$, for $j \in \{1, 2, \ldots, m\}$. It follows from (3) that $\boldsymbol{\pi}(k) = \boldsymbol{\pi}(1)\boldsymbol{\Gamma}^{k-1}$ if the Markov chain is homogeneous, and that

$$P(y_k) = \boldsymbol{\pi}(1)\boldsymbol{\Gamma}^{k-1}\boldsymbol{P}(y_k)\mathbf{1'}. \tag{5}$$

3.5 The Likelihood of Hidden Markov Models

The likelihood of the HMM will be useful when comparing different models. Assume that $\boldsymbol{y}_{1:n} = (y_1, \ldots, y_n)$ is a sample of n observations. The definition of the likelihood for the HMM, L_n, is the joint probability density function, or probability mass function, of $\boldsymbol{y}_{1:n}$. Hence, the likelihood can be defined as

$$L_n = P(y_1, y_2, \ldots, y_n)\,. \tag{6}$$

In general, one can compute the joint distribution as

$$P(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = P(X_1) \prod_{k=2}^{n} P(X_k|X_{k-1}) \prod_{k=1}^{n} P(Y_k|X_k)\,, \tag{7}$$

and the likelihood L_n, by noting that

$$\begin{aligned} L_n &= \sum_{x_1, x_2, \ldots, x_n = 1}^{m} P(X_1 = x_1, \ldots, X_n = x_n, y_1, \ldots, y_n) \\ &= \boldsymbol{\pi}\boldsymbol{P}(y_1)\boldsymbol{\Gamma}\boldsymbol{P}(y_2)\boldsymbol{\Gamma}\boldsymbol{P}(y_3)\cdots\boldsymbol{\Gamma}\boldsymbol{P}(y_n)\mathbf{1'}\,. \end{aligned}$$

3.6 Backward Probabilities

The backward probabilities are defined as

$$P(\boldsymbol{y}_{k+1:n}|X_k = j) = \beta_k(j)\,, \quad j \in \{1, 2, \ldots, m\}\,,\, k \in \{n, n-1, \ldots, 1\}\,.$$

That is, the probability of observing the sequence of random variables $y_{k+1}, \ldots, y_n$, conditional on $X_k = j$. Note that $\beta_n(j) = 1\; \forall\, j \in \{1, 2, \ldots, m\}$.

The backward probabilities can be summarized in $\boldsymbol{\beta}_k$, a row vector of length m, with jth element $\beta_k(j)$. The vector of backward probabilities can for $k \in \{1, 2, \ldots, n\}$ be written as

$$\boldsymbol{\beta}_k^{'} = \boldsymbol{\Gamma P}(y_{k+1})\boldsymbol{\Gamma P}(y_{k+2})\cdots\boldsymbol{\Gamma P}(y_n)\mathbf{1}' = \left(\prod_{s=k+1}^{n} \boldsymbol{\Gamma P}(y_s)\right)\mathbf{1}'\,.$$

Furthermore, the likelihood can also be expressed in terms of the backward probabilities, as

$$L_n = \boldsymbol{\pi P}(y_1)\boldsymbol{\beta}_1'\,.$$

Let us show how to express the backward probabilities in the form of a recursive algorithm. The backward probabilities in $\boldsymbol{\beta}_k$ can be computed recursively as

$$\boldsymbol{\beta}_k^{'} = \boldsymbol{\Gamma P}(y_{k+1})\boldsymbol{\beta}_{k+1}^{'}\,, \quad k \in \{n-1, n-2, \ldots, 1\}\,.$$

Hence, one can compute the backward probabilities from n to 1 recursively by initially setting

$$\boldsymbol{\beta}_n = \mathbf{1}\,,$$

followed by

$$\begin{gathered}
\boldsymbol{\beta}_{n-1}' = \boldsymbol{\Gamma P}(y_n)\boldsymbol{\beta}_n' \\
\vdots \\
\boldsymbol{\beta}_1' = \boldsymbol{\Gamma P}(y_2)\boldsymbol{\beta}_2'\,.
\end{gathered}$$

A proof of the backward probabilities can be found in e.g. Section 4.1.1 in Zucchini and MacDonald (2009). A further note has to be made regarding the implementation of the backward probabilities as a recursive algorithm. The probabilities tend to zero or infinity exponentially fast in the recursions. One solution to this problem is to normalize the backward variables so that they sum to one over j. A thorough discussion on this topic can be found in Sections 3.2.2, 3.4 and 5.1.1.1–2 in Cappé et al. (2005).

4 Bayesian Parameter Estimation

4.1 Bayesian Inference

There are several methods that can be used to estimate the parameters in HMMs. In this paper, we are mainly concerned with estimating the parameters through the Bayesian framework. A thorough explanation of Bayesian inference can be found in e.g. Gelman et al. (2014). Another source of literature with an excellent presentation of Bayesian inference can be found in Held and Bové (2014).

In Bayesian statistics, conclusions about the unknown parameter, θ, are made in terms of probability statements. These probability statements are conditional on the observed value of the data, y, expressed as $P(\theta|y)$, and is referred to as the *posterior distribution*. It contains all the information available about the unknown parameter θ after having observed the data $Y = y$. The posterior distribution is implicitly conditioned on the known values of any covariates, x.

In order to make probability statements about θ given data, y, one must first obtain the joint probability distribution for θ and y. The joint probability distribution, $P(y, \theta)$, can be obtained by the product of the *sampling distribution* or the likelihood, $P(y|\theta)$, and the *prior distribution*, $P(\theta)$. The likelihood function $P(y|\theta)$ expresses the probabilities of data, given the parameter. The prior information $P(\theta)$ expresses the uncertainty about the unknown value θ, before data is observed.

The posterior distribution can be obtained by using Bayes' rule:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \propto P(y|\theta)P(\theta)\,, \tag{8}$$

where $P(y) = \Sigma_\theta P(y|\theta)P(\theta)$ when θ is discrete, i.e. the sum of all possible values of θ. If θ is continuous, $P(y) = \int P(y|\theta)P(\theta)d\theta$. The denominator of the second part of (8) does not depend on θ, and with fixed y can be considered as a constant. Hence, the last part of (8) can be viewed as the *unnormalized posterior density*.

There may be situations where the posterior distribution cannot be computed by analytical methods and computational approximations are needed. *Markov chain Monte Carlo* (MCMC) techniques can be used to simulate dependent draws of θ that converge to the posterior distribution. The sampling is done sequentially, with the distribution of the sampled draws depending on the last value drawn. Thus, the posterior distribution is evaluated by drawing samples from a Markov chain. It is often the case that the first draws in the simulation do not come from the true posterior distribution. Instead, the draws converge after a number of iterations. It is therefore common to discard a part of the initial draws, referred to as a *burn-in* period.

4.2 Gibbs Sampler

The Gibbs sampler is an MCMC technique that can be used to generate random variables from a marginal distribution by iterative sampling from conditional distributions. The method has gained popularity after the seminal paper by Geman and Geman (1984) who studied image-processing models, but the roots of the technique can be traced back to at least Metropolis et al. (1953). An intuitive explanation of the Gibbs sampler can be found in e.g. Casella and George (1992).

The purpose of the Gibbs sampler is to generate random variables from a marginal distribution indirectly, without having to calculate its density. One of the reasons to use the Gibbs sampler technique is to avoid difficult calculations, and instead, replace them with a sequence of easier calculations.

Let us present a simple example to illustrate the use of the technique. Assume that there are two random variables, (X, Y). The aim is to sample from $f(x)$ by sampling from the conditional distributions $f(x|y)$ and $f(y|x)$ instead. Hence, the *Gibbs sequence* of random variables

$$Y^{(0)}, X^{(0)}, Y^{(1)}, X^{(1)}, \dots, Y^{(k)}, X^{(k)}, Y^{(k+1)} \tag{9}$$

is generated by iteratively sampling from

$$\begin{aligned} X^{(k)} &\sim f(x|Y^{(k)} = y^{(k)}) \\ Y^{(k+1)} &\sim f(y|X^{(k)} = x^{(k)}) \,, \end{aligned}$$

where the initial value $Y^{(0)} = y^{(0)}$ is specified. The Gibbs sequence in (9) is commonly referred to as Gibbs sampling, and the distribution of $X^{(k)}$ converges to the marginal distribution $f(x)$ as $k \to \infty$. The proof is omitted here, but an outline of the proof together with a thorough discussion can be found in e.g. Casella and George (1992) and the references therein.

In the Bayesian setting, the mechanism is similar for obtaining the posterior distribution of the unknown parameter vector $\boldsymbol{\vartheta} = (\theta_1, \theta_2, \dots, \theta_n)$, by drawing samples from the full conditional distributions of the parameters of interest.

4.3 Bayesian Derivations

Let us now turn to Bayesian parameter estimation of HMMs. In this section, we present an MCMC method similar to that presented by Rydén (2008), where the Gibbs sampler is used to estimate the parameters of the HMM. An advantage with the Gibbs sampler compared to the EM algorithm is that it does not only produces point estimates, but also gives a measure of the uncertainty about the parameter estimates in the form of the posterior distribution.

The underlying states, $\boldsymbol{X}_{1:n}$, are included as latent data when implementing the Gibbs sampler. In practice, this means that we alternate between sampling model parameters and latent data from their respective full conditional distributions.

Before we present the prior distributions and the full conditional distributions, let us rewrite the joint distribution of the HMM defined in (7) in another form:

$$\begin{aligned} P(\boldsymbol{X}_{1:n}, \boldsymbol{y}_{1:n}) &= \pi_{x_1} \prod_{k=2}^{n} \gamma_{x_{k-1},x_k} \prod_{k=1}^{n} P(y_k|\mu_{x_k}, \sigma^2_{x_k}) \\ &= \pi_{x_1} \prod_{i=1}^{m} \prod_{j=1}^{m} \gamma_{ij}^{n_{ij}} \prod_{j=1}^{m} \prod_{k:X_k=j} P(y_k|\mu_j, \sigma^2_j) \,, \end{aligned}$$

where $n_{ij} = \#\{1 < k \leq n : X_{k-1} = i, X_k = j\}$ is the number of transitions from state i to state j in the latent state sequence. This form of the joint distribution will be useful when deriving the full conditional distributions in the Gibbs sampler.

Moreover, the number of observations in state j can be written as

$$n_j = \sum_{k=1}^{n} I\{X_k = j\}\,,$$

where $I\{\cdot\}$ denotes an indicator function.

4.3.1 Prior Distributions

A prior distribution is specified for each of the parameters in the parameter vector $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\Gamma}, \boldsymbol{\mu}, \boldsymbol{\sigma}^{\mathbf{2}})$, where $\boldsymbol{\pi}$ denotes the initial distribution of the HMM, $\boldsymbol{\Gamma}$ the one-step transition probability matrix, $\boldsymbol{\mu}$ the vector of mean values of the different states, and $\boldsymbol{\sigma}^{\mathbf{2}}$ the vector of the variances of the different states.

Initial Distribution

The initial distribution, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$, is given a *Dirichlet* prior distribution, $\mathrm{Dir}(1, \ldots, 1)$. Thus,

$$(\pi_1, \ldots, \pi_m) \sim \mathrm{Dir}\,(1, \ldots, 1)\,. \tag{10}$$

The Dirichlet distribution is a multivariate generalization of the *Beta* distribution. The probability density function of the Dirichlet distribution $(\pi_1, \ldots, \pi_m) \sim \mathrm{Dir}\,(\alpha_1, \ldots, \alpha_m)$ is defined as

$$P(\boldsymbol{\pi}) = \frac{\Gamma\left(\Sigma_{j=1}^{m}\alpha_j\right)}{\prod_{j=1}^{m}\Gamma(\alpha_j)} \prod_{j=1}^{m} \pi_j^{\alpha_j - 1}\,,$$

where $\pi_1, \pi_2, \ldots, \pi_m \in (0, 1)$, $\Sigma_{j=1}^{m}\pi_j = 1$ and $\alpha_j > 0\ \forall j \in \{1, 2, \ldots, m\}$. The *prior parameters*, commonly referred to as the *hyperparameters*, are specified as $\alpha_1, \alpha_2, \ldots, \alpha_m$. If one has no prior information to favor one state over the other, a symmetric prior distribution where all parameters are equal can be specified, which is done in our case. See e.g. Lin (2016) and references therein for a thorough explanation of the Dirichlet distribution.

Transition Probabilities

Similar to the initial distribution, each row of the transition probability matrix $\boldsymbol{\Gamma}$ is given an independent Dirichlet prior. Hence,

$$(\gamma_{j1}, \ldots, \gamma_{jm}) \sim \mathrm{Dir}\,(1, \ldots, 1)\,, \quad j \in \{1, 2, \ldots, m\}\,. \tag{11}$$

Joint Prior for the Mean and Variance

Derivations of the full conditional distributions are presented in the next subsection. However, some comments must be made regarding the strategy that is used to obtain the full conditional distribution of (μ_j, σ_j^2), for $j \in \{1, 2, \ldots, m\}$. The full conditional distribution of μ_j and σ_j^2 can be obtained using the relation

$$P(\mu_j, \sigma_j^2 | \ldots) \propto \prod_{k:X_k=j} P(y_k|\mu_j, \sigma_j^2) P(\mu_j, \sigma_j^2)\,,$$

where '...' denotes other parameters in the parameter vector $\boldsymbol{\vartheta}$, the hidden Markov chain $\boldsymbol{X}_{1:n}$, and the data $\boldsymbol{y}_{1:n}$. We use this notation in order to express the full conditional distributions throughout this paper, although the variables that affect the full conditional distribution are explicitly expressed. Furthermore, independence is assumed across all j.

Moreover, the joint prior distribution for the jth state can be obtained by using the relation

$$P(\mu_j, \sigma_j^2) = P(\mu_j|\sigma_j^2) P(\sigma_j^2)\,.$$

Assuming that the observations in each state follow a Normal distribution, a commonly used prior for the variance is the *Inverse Gamma distribution:*

$$P(\sigma_j^2) = \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} (\sigma_j^2)^{-(\alpha_j+1)} \exp\left\{-\beta_j/\sigma_j^2\right\} \propto (\sigma_j^2)^{-(\alpha_j+1)} \exp\left\{-\beta_j/\sigma_j^2\right\}\,,$$

which has hyperparameters (α_j, β_j). A convenient reparameterization of the prior distribution gives the scaled inverse-χ^2 distribution with hyperparameters σ_{0j}^2 as scale, and υ_{0j} degrees of freedom. Thus, the prior distribution of σ_j^2 is the distribution of $\sigma_{0j}^2 \upsilon_{0j}/X$, where $X \sim \chi^2_{\upsilon_{0j}}$, i.e. a Chi-square random variable with υ_{0j} degrees of freedom. If

$$\sigma_j^2 \sim \text{Scaled inv-}\chi^2\left(\upsilon_{0j}, \sigma_{0j}^2\right), \text{ then } \sigma_j^2 \sim \text{Inv-Gamma}\left(\frac{\upsilon_{0j}}{2}, \frac{\upsilon_{0j}\sigma_{0j}^2}{2}\right),$$

and hence, the distribution of a Scaled inv-χ^2 distribution with hyperparameters $(\upsilon_{0j}, \sigma_{0j}^2)$ can be written as

$$P(\sigma_j^2) = \frac{(\sigma_{0j}^2 \upsilon_{0j}/2)^{\upsilon_{0j}/2}}{\Gamma(\upsilon_{0j}/2)} (\sigma_j^2)^{-(\upsilon_{0j}/2+1)} \exp\left\{-\frac{\upsilon_{0j}\sigma_{0j}^2}{2\sigma_j^2}\right\} \propto (\sigma_j^2)^{-(\upsilon_{0j}/2+1)} \exp\left\{-\frac{\upsilon_{0j}\sigma_{0j}^2}{2\sigma_j^2}\right\}.$$

The prior distribution can be viewed as providing information equivalent to υ_{0j} observations with an average square deviation σ_{0j}^2. A thorough explanation of the Scaled inverse-χ^2 distribution can be found in Section 2.6 in Gelman et al. (2014).

Moreover, the conditional prior distribution of μ_j given σ_j^2 is a Normal distribution with parametrization

$$\mu_j | \sigma_j^2 \sim N(\mu_{0j}, \sigma_j^2/\kappa_{0j})\,.$$

Thus, the conditional prior density can be written as

$$P(\mu_j|\sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2/\kappa_{0j}}} \exp\left\{-\frac{(\mu_j - \mu_{0j})^2}{2\sigma_j^2/\kappa_{0j}}\right\} \propto \frac{1}{\sigma_j} \exp\left\{-\frac{(\mu_j - \mu_{0j})^2}{2\sigma_j^2/\kappa_{0j}}\right\} .$$

The joint prior density $P(\mu_j, \sigma_j^2)$ can be obtained by multiplying the unconditional prior variance $P(\sigma_j^2)$ and the conditional prior mean $P(\mu_j|\sigma_j^2)$:

$$\begin{aligned} P(\mu_j, \sigma_j^2) &= P(\mu_j|\sigma_j^2)P(\sigma_j^2) \\ &\propto \sigma_j^{-1} \exp\left\{-\frac{(\mu_j - \mu_{0j})^2}{2\sigma_j^2/\kappa_{0j}}\right\} \times (\sigma_j^2)^{-(\upsilon_{0j}/2+1)} \exp\left\{-\frac{\upsilon_{0j}\sigma_{0j}^2}{2\sigma_j^2}\right\} \\ &= \sigma_j^{-1}(\sigma_j^2)^{-(\upsilon_{0j}/2+1)} \exp\left\{-\frac{1}{2\sigma_j^2}\left[\upsilon_{0j}\sigma_{0j}^2 + \kappa_{0j}(\mu_{0j} - \mu_j)^2\right]\right\} . \end{aligned} \tag{12}$$

This distribution can be labeled as Scaled N-inv-$\chi^2(\mu_{0j}, \sigma_{0j}^2/\kappa_{0j}; \upsilon_{0j}, \sigma_{0j}^2)$, a four parameter model. An interpretation of the parameters is that the first two parameters represent the location and scale of μ, and the last two represent the degrees of freedom and scale of σ^2. The Scaled N-inv-χ^2-distribution can be viewed as a reparametrization of the four parameter *Normal-inverse-gamma distribution*, see Appendix A.1 for its probability density function.

4.3.2 Derivation of Full Conditional Distributions

Before the derivations of the full conditional distributions are presented, the term *conjugacy* must be introduced. A prior is conjugate to the likelihood if the prior and posterior belong to the same distributional family.

The formal definition of conjugacy in accordance with Gelman et al. (2014) is defined as: If $\mathcal{F}$ is a class of sampling distributions $P(y|\theta)$, and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$P(\theta|y) \in \mathcal{P} \text{ for all } P(\cdot|\theta) \in \mathcal{F} \text{ and } P(\cdot) \in \mathcal{P} .$$

Full Conditional Distribution of the Initial Distribution

The Dirichlet distribution is conjugate to the *Multinomial distribution*, which is a generalization of the *Binomial* distribution. The sampling distribution of the initial probability is assumed to follow a Multinomial distribution, which has the probability mass function

$$P(n_1, \ldots, n_m|n, \pi_1, \ldots, \pi_m) = \frac{n!}{n_1! \cdot \ldots \cdot n_m!} \prod_{j=1}^{m} \pi_j^{n_j} ,$$

where $n_j \in \{0, 1, \ldots, n\}$, $\Sigma n_j = n$ and $\pi_1, \ldots, \pi_m$ are the event probabilities. Here, $n_j = \#\{1 \leq k \leq n : X_k = j\}$ is the number of visits to state j in the latent state sequence.

Moreover, the full conditional distribution of the initial distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_m)$ can be written as

$$\begin{aligned} P(\boldsymbol{\pi}|\boldsymbol{X}_{1:n}, \ldots) &\propto P(n_1, \ldots, n_m | n, \pi_1, \ldots, \pi_m) P(\pi_1, \ldots, \pi_m) \\ &\propto \left(\frac{n!}{n_1! \cdot \ldots \cdot n_m!} \prod_{j=1}^{m} \pi_j^{n_j} \right) \left(\frac{\Gamma\left(\Sigma_{j=1}^{m} \alpha_j\right)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \pi_j^{\alpha_j - 1} \right) \\ &\propto \prod_{j=1}^{m} \pi_j^{n_j + \alpha_j - 1} \sim \text{Dir}\left(n_1 + \alpha_1, \ldots, n_m + \alpha_m\right). \end{aligned}$$

At the first time point, only one state can govern the distribution of the initial volume of the solder paste droplet. Thus, $n = 1$, and, due to the prior specification in (10), the full conditional distribution is given by

$$(\pi_1, \ldots, \pi_m)|\boldsymbol{X}_{1:n}, \ldots \sim \text{Dir}\ \left(I\{X_1 = 1\} + 1, \ldots, I\{X_1 = m\} + 1\right). \tag{13}$$

Full Conditional Distribution of the Transition Probability Matrix

Similar to the initial probabilities, the sampling distribution of each row in the transition probability matrix is assumed to follow a Multinomial distribution. The full conditional distribution of the rows in the transition probability matrix is therefore given by

$$(\gamma_{i1}, \ldots, \gamma_{im})|\boldsymbol{X}_{1:n}, \ldots \sim \text{Dir}\left(n_{i1} + 1, n_{i2} + 1, \ldots, n_{im} + 1\right). \tag{14}$$

Recall that n_{ij} is the number of transitions from state i to state j in the latent state sequence. Furthermore, conditional independence is assumed across the rows in the transition probability matrix.

Full Conditional Distribution for the Mean and the Variance

Recall that we defined the number of visits to state j in the latent state sequence as n_j, with conditional independence across $j \in \{1, \ldots, m\}$. Furthermore, the sample mean of state j is defined as $\bar{y}_j = \Sigma_{k:\, X_k = j}\, y_k / n_j$. Since the data is assumed to follow a Normal distribution in each state, the joint probability of the random variables in state j, for $j \in \{1, \ldots, m\}$, can be written as:

$$\begin{aligned} \prod_{k:X_k=j} P(y_k|\mu_j, \sigma_j^2) &= \prod_{k:X_k=j} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{(y_k - \mu_j)^2}{2\sigma_j^2} \right\} \\ &\propto (\sigma_j^2)^{-n_j/2} \exp\left\{ -\frac{1}{2\sigma_j^2} \left[(n_j - 1)s_j^2 + n_j(\bar{y}_j - \mu_j)^2 \right] \right\}, \end{aligned}$$

where $s_j^2 = \frac{1}{n_j - 1} \Sigma_{k:X_k=j} (y_k - \bar{y}_j)^2$. The full expression of the joint distribution is given in Appendix A.2.

The full conditional distribution of the jth state can thus be written as

$$\begin{aligned} P(\mu_j, \sigma_j^2 | \boldsymbol{X}_{1:n}, \boldsymbol{y}_{1:n}, \ldots) \propto \sigma_j^{-1} (\sigma_j^2)^{-(\upsilon_{0j}/2+1)} \exp\left\{ -\frac{1}{2\sigma_j^2} \left[\upsilon_{0j}\sigma_{0j}^2 + \kappa_{0j}(\mu_{0j} - \mu_j)^2 \right] \right\} \times \\ \times (\sigma_j^2)^{-n_j/2} \exp\left\{ -\frac{1}{2\sigma_j^2} \left[(n_j - 1)s_j^2 + n_j(\bar{y}_j - \mu_j)^2 \right] \right\} \\ \propto \text{Scaled N-inv-}\chi^2(\tilde{\mu}_j, \tilde{\sigma}_j^2/\tilde{\kappa}_j; \tilde{\upsilon}_j, \tilde{\sigma}_j^2)\,, \end{aligned} \tag{15}$$

where

$$\begin{aligned} \tilde{\mu}_j &= \frac{\kappa_{0j}}{\kappa_{0j} + n_j}\mu_{0j} + \frac{n_j}{\kappa_{0j} + n_j}\bar{y}_j \\ \tilde{\kappa}_j &= \kappa_{0j} + n_j \\ \tilde{\upsilon}_j &= \upsilon_{0j} + n_j \\ \tilde{\upsilon}_j\tilde{\sigma}_j^2 &= \upsilon_{0j}\sigma_{0j}^2 + (n_j - 1)s_j^2 + \frac{\kappa_{0j}n_j}{\kappa_{0j} + n_j}(\bar{y}_j - \mu_{0j})^2 \,. \end{aligned}$$

The parameters of the full conditional distribution combine the prior information and the information obtained from the data. A thorough explanation of the full conditional distribution can be found in Section 3.3 in Gelman et al. (2014). Furthermore, a proper derivation of the joint full conditional distribution is given in Appendix A.3.

The full conditional density of μ_j given $\sigma_j^2, \boldsymbol{y}_{1:n}$, $\boldsymbol{X}_{1:n}$, and the other parameters is proportional to the joint posterior density in (15), with σ_j^2 held constant, such that

$$\mu_j | \sigma_j^2, \boldsymbol{X}_{1:n}, \boldsymbol{y}_{1:n}, \ldots \sim N\left(\tilde{\mu}_j, \frac{\sigma_j^2}{\tilde{\kappa}_j} \right) \,. \tag{16}$$

Moreover, the marginal full conditional distribution of σ_j^2 given $\boldsymbol{y}_{1:n}$, $\boldsymbol{X}_{1:n}$ and the other parameters is Scaled inverse-χ^2:

$$\sigma_j^2 | \boldsymbol{X}_{1:n}, \boldsymbol{y}_{1:n}, \ldots \sim \text{Scaled inv-}\chi^2 \left(\tilde{\upsilon}_j, \tilde{\sigma}_j^2 \right) \,. \tag{17}$$

One can sample from the full conditional distribution, $P(\mu_j, \sigma_j^2 | \boldsymbol{X}_{1:n}, \boldsymbol{y}_{1:n}, \ldots)$, by following these two steps:

1. Draw σ_j^2 from the full conditional distribution in (17).

2. Draw μ_j from the full conditional distribution in (16), using the simulated value of σ_j^2 from (17).

In our simulation, the vector of hyperparameters, $\boldsymbol{\kappa}_0, \boldsymbol{\upsilon}_0, \boldsymbol{\sigma}_0$, are given the prior $\mathbf{1}$. The vector of mean hyperparameters, $\boldsymbol{\mu}_0$, are evenly spread across the range of data.

4.4 Sampling the Hidden Markov Chain

The Gibbs sampler alternates between updating the parameters in the parameter vector $\boldsymbol{\vartheta}$ and the hidden Markov chain $\boldsymbol{X}_{1:n}$. The hidden Markov chain can be updated using the following steps:

1. Compute the probability $P(X_1 = j|\boldsymbol{y}_{1:n}, \boldsymbol{\vartheta})$, for $j \in \{1, \ldots, m\}$.
2. Sample X_1.
3. For $k = 2, \ldots, n$,
   (a) Compute the probability $P(X_k = j|\boldsymbol{y}_{1:n}, \boldsymbol{X}_{1:k-1}, \boldsymbol{\vartheta})$, for $j \in \{1, \ldots, m\}$.
   (b) Sample X_k.

Sampling the Initial State

The probability that $X_1 = j$, for $j \in \{1, 2, \ldots, m\}$, given the parameters and the data, is proportional to:

$$\begin{aligned} P(X_1 = j|\boldsymbol{y}_{1:n}, \boldsymbol{\vartheta}) &\propto P(X_1 = j) P(y_1|\mu_j, \sigma_j^2)\, p(\boldsymbol{y}_{2:n}|X_1 = j) \\ &= \pi_j f_j(y_1) \beta_1(j)\,. \end{aligned} \tag{18}$$

Since the probabilities are only proportional in j, they need to be normalized in order to obtain the correct probabilities. The derivation for computing the probabilities is given in Appendix A.4.1.

Updating the States Conditional on the Previously Sampled States

The probability that $X_k = j$, conditional on the previously sampled values of the chain, $\boldsymbol{X}_{1:k-1}$, and data, $\boldsymbol{y}_{1:n}$, is a non-homogeneous Markov chain with transition probabilities

$$\begin{aligned} P(X_k = j|\boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:n}, \boldsymbol{\vartheta}) &\propto P(X_k = j|X_{k-1} = i) P(y_k|X_k = j) P(\boldsymbol{y}_{k+1:n}|X_k = j) \\ &= \gamma_{ij} f_j(y_k) \beta_k(j)\,. \end{aligned} \tag{19}$$

Similar to the procedure of sampling the initial state, the probabilities have to be normalized at each iteration, before the new state is sampled. Note that the densities of the partial data correspond to the backward probabilities. Since the backward probabilities have to be obtained, the method for simulating the latent Markov chain conditional on data and the parameter vector is called *backward recursion forward sampling*. A derivation of the process of computing the probabilities is given in Appendix A.4.2.

4.5 Simulation Method

The algorithm can be initialized by guessing the initial states of the hidden Markov chain.

The following implementation of the Gibbs sampler is used:

1. First step

   (a) Update σ_j^2 by drawing independently from (17), for $j \in \{1, \ldots, m\}$. Update μ_j by drawing independently from (16), for $j \in \{1, \ldots, m\}$.

   (b) Update $\boldsymbol{\pi}$ by drawing $(\pi_1, \ldots, \pi_m)$ from (13).

   (c) Update $\boldsymbol{\Gamma}$ by drawing $(\gamma_{j1}, \ldots, \gamma_{im})$ from (14), independently for $j \in \{1, 2, \ldots, m\}$.

2. Second step

   (a) Update $\{X_k\}_{k=1}^n$ by drawing X_1 from (18) and then X_k from (19), for $k \in \{2, 3, \ldots, n\}$.

One sequence of these steps (first and second step) is referred to as a *sweep* of the Gibbs sampler.

4.6 Model Selection

Let us now briefly discuss likelihood-based methods that can be used to deal with model uncertainty in HMMs. A thorough discussion on this topic can be found in e.g. Sections 4.4.1-2 in Frühwirth-Schnatter (2006), or in Sections 2.1-2.2.1 in Celeux et al. (2018).

We are interested in finding the true number of states, m, when choosing among several models. *Information criteria* are based on penalizing the log-likelihood function, in this case, the log-likelihood of the HMM. A model with m states is defined as $\mathcal{M}_m$. The log-likelihood, $\ell(\boldsymbol{\vartheta}; m)$, is defined as the natural logarithm of the likelihood of a model $\mathcal{M}_m$, where the likelihood of the HMM is defined in (6). The penalty is proportional to the number of free parameters in $\mathcal{M}_m$, denoted υ_m. There are several information criteria that can be used for selecting the proper model, and perhaps the two most famous being the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978).

The aim is to minimize AIC(m), which is defined as

$$\text{AIC}(m) = -2\ell(\hat{\boldsymbol{\vartheta}}_m; m) + 2\upsilon_m\,,$$

where $\hat{\boldsymbol{\vartheta}}$ is the *maximum likelihood* estimator of the model $\mathcal{M}_m$. Similarly, the aim is to minimize BIC(m), which is defined as

$$\text{BIC}(m) = -2\ell(\hat{\boldsymbol{\vartheta}}_m; m) + \upsilon_m \ln(n)\,,$$

where n is the total number of observations. Since the Gibbs sampler is used in our simulation, the maximum likelihood is approximated by taking the maximum over the likelihoods that are simulated during the sweeps. Furthermore, the number of free parameters, υ_m, can be computed by noting that the parameter vector consists of $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\Gamma}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$. The initial probabilities $\boldsymbol{\pi}$ contain $m-1$ number of free parameters, where m denotes the total number of states. Similarly, the transition probability matrix $\boldsymbol{\Gamma}$ contains $m^2 - m$ number of free parameters, and $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ contain m number of free parameters, respectively. Thus, the number of free parameters in the model are $m^2 + 2m - 1$.

Moreover, let us briefly mention how to compute the log-likelihood by using the backward probabilities. Recall that the backward probabilities, $\boldsymbol{\beta}_k$, for $k \in \{n, n-1, \ldots, 1\}$ are normalized during the recursions. In order to obtain the log-likelihood, the normalizing constant, C_k, is saved to a vector during the recursions, for $k \in \{n, n-1, \ldots, 1\}$. Furthermore, C_0 is defined as $\boldsymbol{\pi}\boldsymbol{P}(y_1)\tilde{\boldsymbol{\beta}}_1'$, where $\tilde{\boldsymbol{\beta}}_1$ is a vector with normalized backward probabilities. The log-likelihood can be computed by noting that

$$\ln L_n = \sum_{k=0}^{n} \ln C_k \,.$$

5 Results

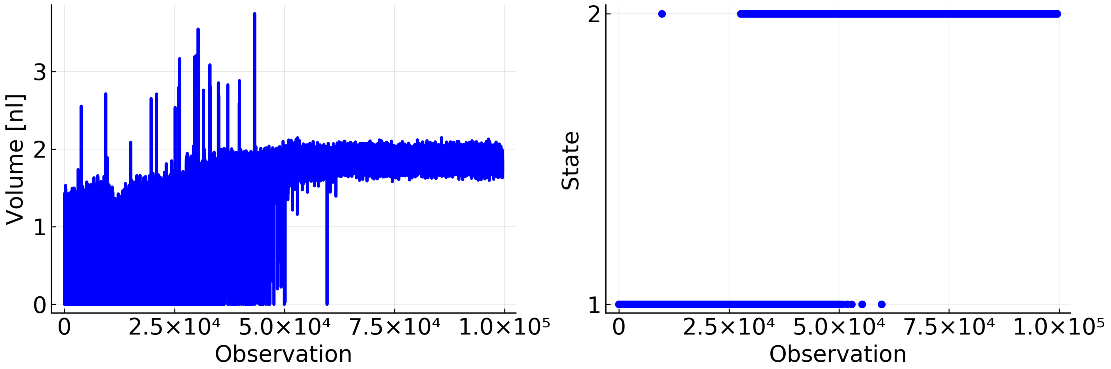
The Gibbs sampler is used to sample from the full conditional distributions. The two different types of data sets described in Sections 2.1.1 and 2.1.2 are used in the simulations. The first application of the Gibbs sampler is to estimate ξ, the number of solder paste droplets that are jetted before reaching the system equilibrium. The second application of the Gibbs sampler is to estimate the mean and standard deviation of the different states of the system, assuming that the droplets are jetted according to the system equilibrium.

Extensive work has been carried out by Raftery and Lewis (1992) on the number of iterations that are required in the Gibbs sampler, in order to estimate a posterior quantile within a margin of error with a certain probability. The results suggest that reasonable accuracy can often be achieved with 5 000 iterations or less. The number of iterations used in the Gibbs sampler in this paper is 11 000, where the first 1 000 draws are discarded as burn-in values. Thus, the analysis is carried out on 10 000 draws from the full conditional distributions.

5.1 Simulation of the Initiation Period

The data set with 99 360 observations described in Section 2.1.1 is used in this simulation. Recall that the hidden states $\boldsymbol{X}_{1:n}$ are sampled during each sweep of the Gibbs sampler. We assume that there are two states that govern the distribution of the volume. State 1 represents the initiation period and state 2 represents the system equilibrium. The aim is to estimate ξ, the number of droplets that are jetted before reaching the system equilibrium. The Gibbs sampler not only produces a point estimate of this quantity, but also allows for a measure of the uncertainty of the estimate, in the form of a posterior distribution of the last time point that the system is in the initiation period.

Figure 10a shows a plot of the volumes of the droplets in the data set, and Figure 10b shows a plot of the sampled states during one of the sweeps in our simulation. The number of droplets that are jetted before reaching the system equilibrium during the sweep is approximately 60 000 droplets. The last time point that the system is in the initiation period is recorded for each sweep in the Gibbs sampler, which results in a posterior distribution of ξ.



(a)

(b)

Figure 10: *Illustration of the process of computing ξ, where (a) represents the measurements on the volume of the droplets and (b) represents the sampled states during one of the Gibbs sweeps.*

The posterior distribution of ξ is shown in Figure 11. In general, the *skewness* can be used as a measure of the distributional asymmetry. A skewness close to zero indicates that the histogram is symmetric around its mean. Since the distribution of ξ is skewed, with skewness of 1.758, the median will serve as our point estimate. The median is 59 700 droplets, and the histogram shows that the majority of the computed ξ's in the MCMC simulation are concentrated around the time point 60 000.

One could also use the posterior draws to compute a credible interval for the last time point that the system visits state 1. The mean based on the Gibbs sampling output is 65 800 droplets. Furthermore, a 95% credible interval of ξ is given by 59 700 $\leq \xi \leq$ 96 300. Hence, the probability that ξ lies between 59 700 and 96 300 is 0.95. The posterior quantiles can also be used to compute an upper bound for ξ. In our simulation, the 0.95 quantile corresponds to 93 300 droplets.

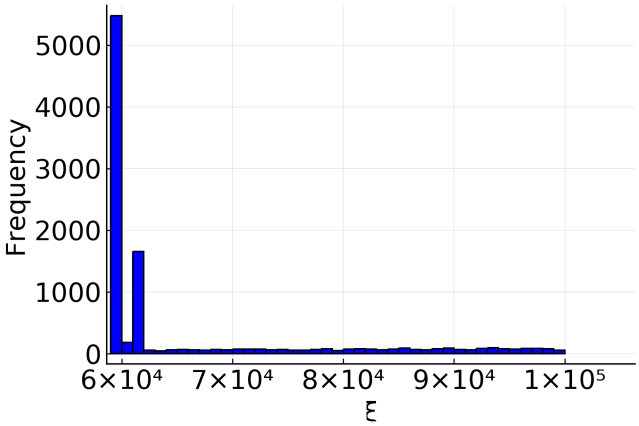


Figure 11: *The histogram shows the distribution of ξ, i.e. the last time point that the system is in the initiation period. The initial sample contained 11 000 draws, and the first 1 000 draws were discarded as burn-in values.*

5.2 Simulation of the Parameters During System Equilibrium

The data set with 49 000 observations described in Section 2.1.2 is used in this simulation. The aim is to estimate the mean and the standard deviation of the different states, together with a measure of the uncertainty of the estimates, in the form of a 95% credible interval for the parameters. Furthermore, the estimation of the transition probability matrix $\mathbf{\Gamma}$ is also considered.

In the previous simulation, we assumed that there were only two different states present during the jetting. The system was either in state 1 or state 2, i.e. either in the initiation period or in the system equilibrium. However, specifying the number of hidden states that govern the distribution of the volume may not always be as trivial. Analyzing Figure 7a visually, there seem to be several states present during the jetting. A scatter plot of the same data set is presented in Figure 12, in order to give a better overview of the data.

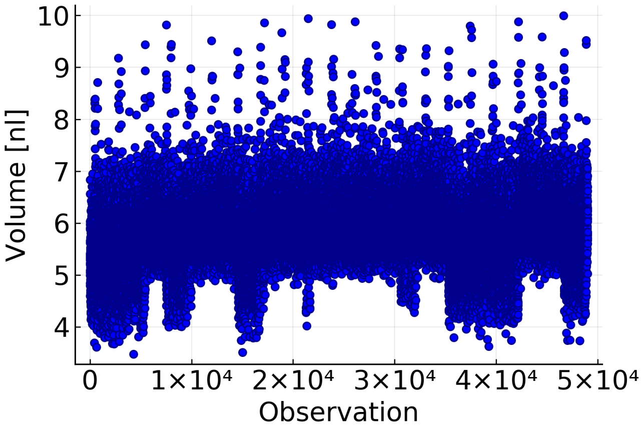


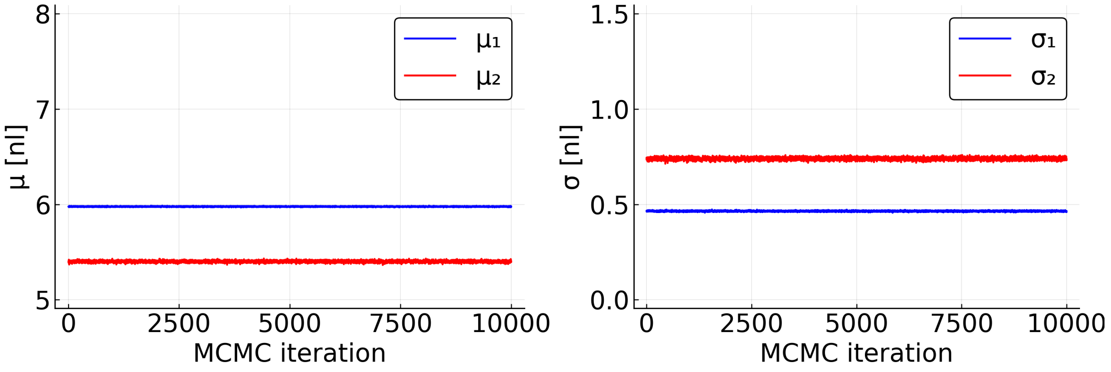
Figure 12: *Plot of the volume of the droplets during system equilibrium. The reference volume is 5.8 nl, and the data set is the same as in Section 2.1.2, with 49 000 observations.*

We assume that there are three different states present during the system equilibrium in this data set, since there seem to be roughly three overlapping clusters of data.

Moreover, we are also interested in comparing the model with three states with other models that may seem probable. Parameter estimation with two different states is therefore also carried out using the same data set, followed by a model comparison through AIC and BIC.

5.2.1 Simulation with Two Different States

Two states were specified in this simulation. Trace plots of the sampled parameters in the MCMC simulation are shown in Figure 13a and Figure 13b. Furthermore, the posterior estimates of the parameters in the two different states are given in Table 1 and Table 2.



(a)

(b)

Figure 13: *MCMC simulation using the Gibbs sampler. The initial simulation contained 11 000 iterations and the first 1 000 observations were discarded as burn-in values.*

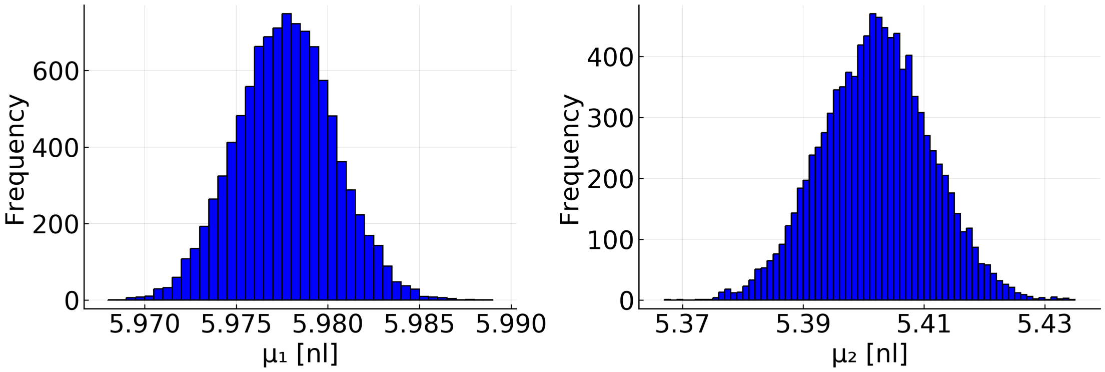
Table 1: *Posterior estimates of the different means in the different states. The mean represents the estimate of μ_j, for $j \in \{1, 2\}$.*

μ	**Mean**	**0.025 Quantile**	**0.975 Quantile**
State 1	5.977	5.972	5.982
State 2	5.402	5.384	5.419

Table 2: *Posterior estimates of the different standard deviations in the different states. The mean represents the estimate of σ_j, for $j \in \{1, 2\}$.*

σ	**Mean**	**0.025 Quantile**	**0.975 Quantile**
State 1	0.465	0.461	0.469
State 2	0.740	0.728	0.752

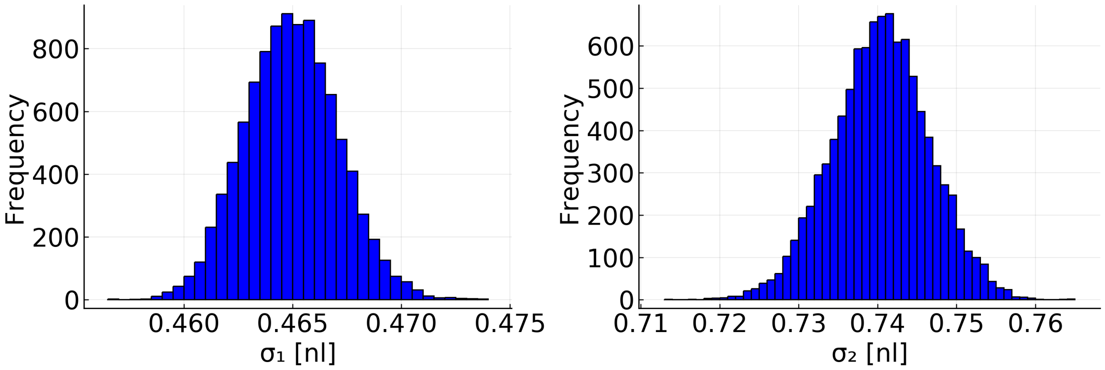
The histograms presented in Figure 14a and Figure 14b give a picture of the estimated posterior distributions of μ_j, for $j \in \{1, 2\}$.



(a) **(b)**

Figure 14: *Histogram of the posterior distributions of the two different means. The initial simulation contained 11 000 iterations, of which the first 1 000 draws were discarded as burn-in values.*

Similarly, the histograms presented in Figure 15a, and Figure 15b give a picture of the estimated posterior distributions of σ_j, for $j \in \{1, 2\}$.



(a) **(b)**

Figure 15: *Histogram of the posterior distributions of the two different standard deviations. The initial simulation contained 11 000 iterations, of which the first 1 000 draws were discarded as burn-in values.*

Moreover, the transition probability matrix $\mathbf{\Gamma}$, which is estimated from 10 000 draws, is given by:

$$\mathbf{\Gamma} = \begin{pmatrix} 0.990 & 0.010 \\ 0.029 & 0.971 \end{pmatrix} .$$

The estimate of the mean in state 1 is closer to the reference, compared to the estimate of the mean in state 2. Furthermore, the estimated standard deviation in state 1 is smaller than the estimated standard deviation in state 2. Given the number of specified states in this simulation, an interpretation of the results is that the droplets that are generated by state 1 are concentrated around 6 nl, and jetted with much smaller variance, compared to the droplets that are generated by state 2, which are jetted around 5.4 nl.

5.2.2 Simulation with Three Different States

Three states were specified during this simulation. Trace plots of the sampled parameters in the MCMC simulation are shown in Figure 16a and Figure 16b. Furthermore, the posterior estimates of the parameters in the two different states are given in Table 3 and Table 4.

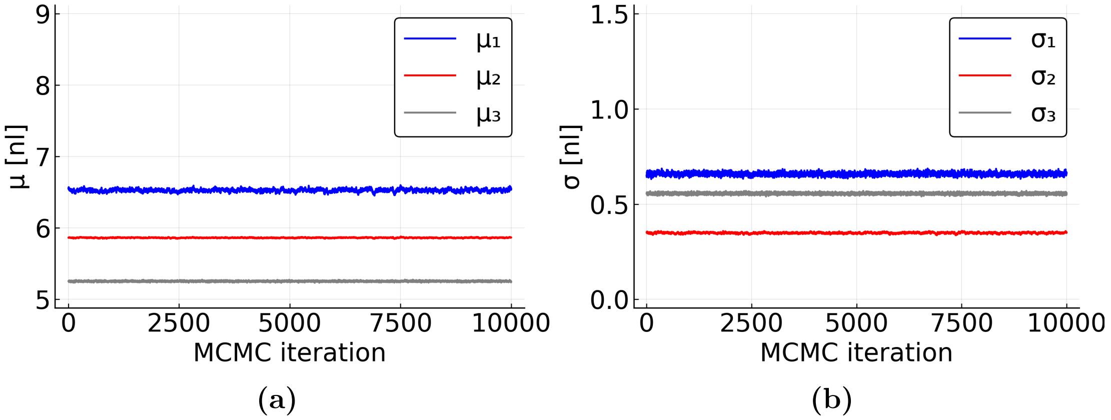


Figure 16: *MCMC simulation using the Gibbs sampler. The initial simulation contained 11 000 iterations and the first 1 000 observations were discarded as burn-in values.*

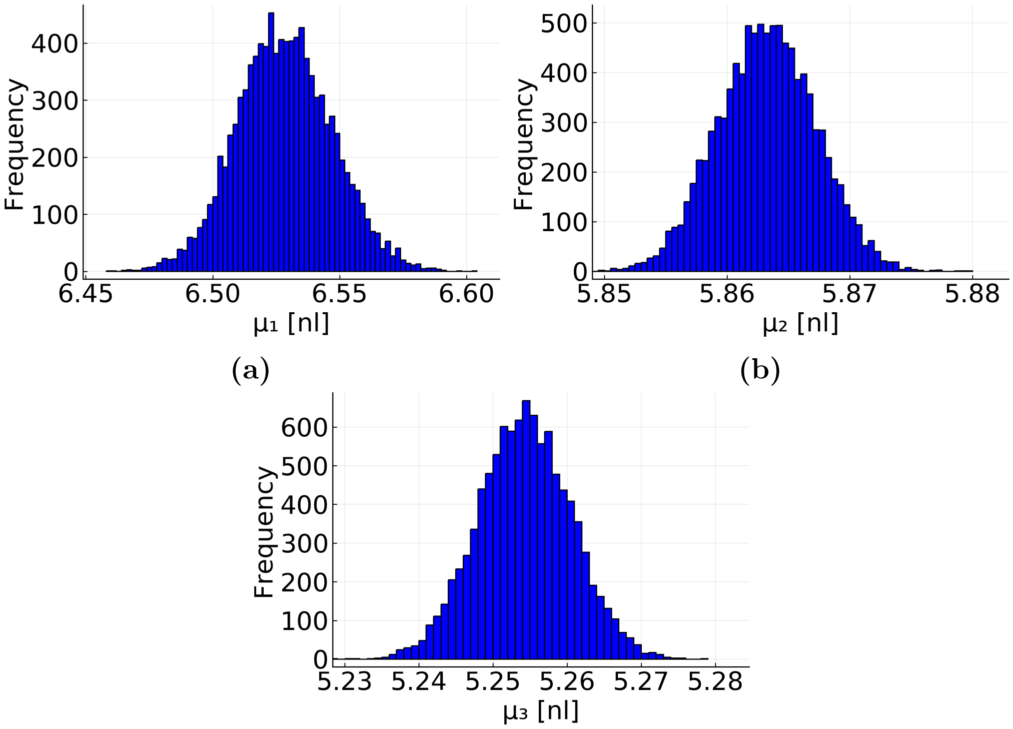
Table 3: *Posterior estimates of the different means in the different states. The mean represents the estimate of μ_j, for $j \in \{1, 2, 3\}$.*

μ	**Mean**	**0.025 Quantile**	**0.975 Quantile**
State 1	6.528	6.492	6.565
State 2	5.865	5.855	5.870
State 3	5.254	5.242	5.266

Table 4: *Posterior estimates of the different standard deviations in the different states. The mean represents the estimate of σ_j, for $j \in \{1, 2, 3\}$.*

σ	**Mean**	**0.025 Quantile**	**0.975 Quantile**
State 1	0.659	0.644	0.675
State 2	0.349	0.343	0.355
State 3	0.556	0.548	0.564

The histograms presented in Figure 17a, Figure 17b and Figure 17c give a picture of the estimated posterior distributions of μ_j, for $j \in \{1, 2, 3\}$.



(a) **(b)**

(c)

Figure 17: *Histogram of the posterior distributions of the different means, where 10 000 posterior draws were sampled.*

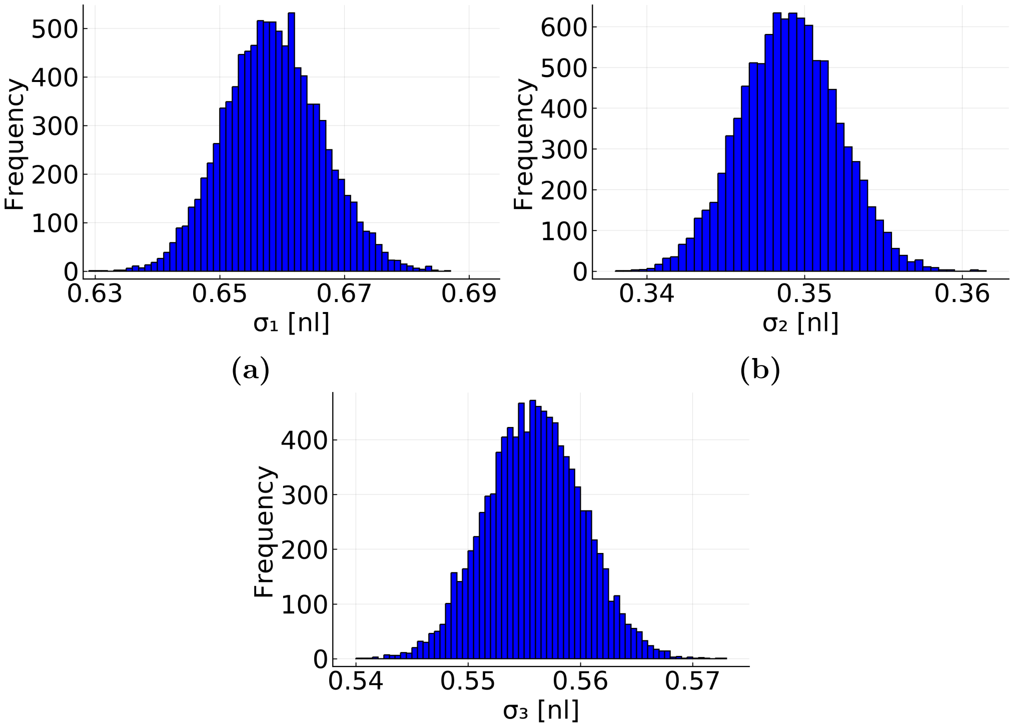
Similarly, the histograms presented in Figure 18a, Figure 18b and Figure 18c give a picture of the estimated posterior distributions of σ_j, for $j \in \{1, 2, 3\}$.

Moreover, the transition probability matrix $\mathbf{\Gamma}$, which is estimated from 10 000 draws, is given by

$$\mathbf{\Gamma} = \begin{pmatrix} 0.420 & 0.559 & 0.021 \\ 0.140 & 0.850 & 0.010 \\ 0.019 & 0.017 & 0.964 \end{pmatrix}.$$

Our parameter estimates show that the solder paste deposits are jetted according to the reference when state 2 governs the distribution of the solder paste droplets. The estimated mean in the second state is 5.86 nl, which coincides with the reference volume, 5.85 nl. The estimated standard deviation in the second state is relatively small, 0.349 nl, compared to the estimated standard deviations in the other two states.

Given the number of specified states in this simulation, an interpretation of the results is that the sudden regime changes at the top and the bottom parts of Figure 12 are caused by state 1 and state 3, respectively. Furthermore, droplets that are generated by state 2 are jetted according to the reference, with a relatively small standard deviation.



(a)

(b)

(c)

Figure 18: *Histogram of the posterior distributions of the different standard deviations, where 10 000 posterior draws were sampled.*

5.2.3 Comparison of the Models

Let us now perform a model comparison of several models. $\mathcal{M}_2$ and $\mathcal{M}_3$ represent the models with two and three states, respectively. Furthermore, the HMMs are also compared to the univariate model, i.e. a simple model with one state, denoted by $\mathcal{M}_1$. Table 5 summarizes the AIC and BIC results for the different models.

Table 5: *Model comparison between the different models, where AIC and BIC are used.*

Model	AIC	BIC
$\mathcal{M}_1$	89 477.06	89 516.26
$\mathcal{M}_2$	79 213.13	79 275.16
$\mathcal{M}_3$	72 808.12	72 931.31

The results from the information criteria imply that the third model $\mathcal{M}_3$, with three different states is favorable when comparing it to the univariate model $\mathcal{M}_1$, and the HMM with two states, $\mathcal{M}_2$.

6 Discussion

6.1 Periodicity During System Equilibrium

One assumption that is made about the data is that the underlying process that governs the distribution of the droplets is a homogeneous Markov chain. However, there appears to be a pattern where the volume temporarily increases with a certain regularity in Figure 12. The spikes in data seem to appear in an interval of 2 000 droplets, and the reason for this periodicity during the jetting in unknown.

If this periodic behavior appears in two data sets, where the aim is to compare the parameters within the different states, a strategy to solve this problem could be to treat the periodic data points as missing values in the Gibbs sampler.

6.2 Assumption of Conditional Independence

Another assumption that has been made about the data is that the volumes of the droplets, $\{Y_k\}$, are conditionally independent given the states, $\{X_k\}$. However, the periodicity due to the deviation of the first droplet in a new strip violates this assumption. Similar to the periodicity described in the previous subsection, one could address this problem by including the first droplets in a new strip as missing values in the Gibbs sampler.

6.3 Skewed Posterior Distribution

The estimate of ξ, i.e. the number of droplets that are jetted before reaching the reference volume, is presented in Section 5.1. The posterior distribution of ξ is skewed. This is because the states $\boldsymbol{X}_{1:n}$ are sampled during each sweep of the Gibbs sampler. There will always be a small probability of sampling a state that does not govern the distribution of the volume.

One could address this problem by imposing a condition when computing ξ during each sweep. For instance, in order to include a time point k as the computed value of ξ during one sweep, one must also observe the initiation period during the previous q consecutive time points, where q is an arbitrary number chosen by the experimenter. As an example, assume that k is the last time point that the system is in the initiation period, which is governed by the state X_k. k is included as the computed value of ξ during the sweep if the states $X_{k-1}, \dots X_{k-q}$ also belong to the initiation period, for $k > q$.

In summary, it seems that it is possible to use the Gibbs sampler to estimate ξ when analyzing the volume of the solder paste droplets in jet printing. Estimating ξ every time an ejector is replaced can thus lead to an estimate of the number of droplets on average that one should jet before reaching the system equilibrium.

6.4 Parameter Estimation During System Equilibrium

Something that is noteworthy is that the variation between the MCMC draws is very small in the Gibbs sampler. Furthermore, none of the trace plots seem to drift in either direction, which implies that the burn-in period does not need to be extended.

6.5 Continuous State Space

Throughout this paper, we have assumed that the states $\{X_k\}$ take values in some finite set, S, called the state space. This assumption seems reasonable when analyzing the volume of the droplets during the system equilibrium, due to the sudden regime changes in data. The data presented in Section 2.1.1 seems different however. There is a continued increase in the upper edge of the curve during the first half of the plot in Figure 6. One could argue that the state space, in this case, is continuous and instead use state space models. Even if the state space may be continuous, the advantage of using HMMs, in this case, is that we can approximate the initiation period as a discrete state. This allows for an estimate of ξ, i.e. the number of droplets that are jetted before reaching the system equilibrium.

6.6 Model Comparison

Although the model with three different states is favorable when using AIC and BIC, no measure of the model uncertainty is given using the information criteria.

Moreover, the periodic data during the system equilibrium has to be dealt with before we can make any final conclusions about the models. AIC and BIC may have to be computed based on the new data set, if the periodic data points are treated as missing values in the Gibbs sampler.

6.7 Normality of the Data Within the Different States

Throughout this paper, the distribution of the volume of the solder paste deposit, Y_k, conditional on the current state of the system, X_k, is assumed to be Normal with a certain mean and a certain variance. However, no attempt to verify this assumption has been made.

One approach that can be used to verify the normality of the data within the different states is to rely on the *central limit theorem* (CLT). Informally, one could argue that the difference between the draws in the MCMC simulation is relatively small in our case. Thus, the normality can be evaluated by using one of the Gibbs sweeps. If the observations within the different states indeed follow a Normal distribution, one could standardize the observations by using the sampled means and the sampled variances of the different states during the sweep. Hence, the distribution of the standardized data should approximately follow a standard Normal distribution $\sim N(0,1)$ if the observations are normally distributed within the different states.

7 Conclusions

In this paper, we have presented a Bayesian HMM that has been used to analyze two data sets that contain measurements of the volume of individual solder paste droplets. The last time point that the system is in the initiation period, ξ, has been estimated. Furthermore, the parameters within the different states during the system equilibrium has been estimated, followed by a model comparison of three different models.

The estimate of ξ seems reasonable when comparing the results to the plot of the volume in Figure 6. Furthermore, a condition when computing ξ during the Gibbs sweeps has been introduced in order to reduce the risk of overestimating ξ during the sweeps.

Moreover, the parameter estimates together with the posterior distribution of the different parameters during the system equilibrium can be used to analyze the quality of the solder paste droplets in product development. The model with three different states is favorable compared to the univariate model and the model with two different states when analyzing AIC and BIC from the simulations. However, another model comparison may have to be carried out if the data points that appear to be periodic during the system equilibrium are treated as missing values in the Gibbs sampler.

Looking back at the purpose of this work, a recurring theme throughout this paper has been to use HMMs for product development purposes. The method can be used as a powerful tool to gain more information about systems with regime changing properties. Some areas where HMMs can be used in product development have been touched upon in the paper. However, one should evaluate whether all assumptions about the data are met. We have briefly touched upon the cases where the various assumptions may not have been fulfilled. A suggestion for further research is therefore to try to handle the periodic data that arises during the jetting process. Another potential area for further research is to evaluate the normality of the data within the different states. Furthermore, one could also analyze if there are other distributions that better describe the distribution of the volume of the deposited fluids within the different states.

References

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Baum, L. E. and Petrie, T. (1966), 'Statistical Inference for Probabilistic Functions of Finite State Markov Chains', *The Annals of Mathematical Statistics* **37**(6), 1554–1563.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970), 'A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains', *The Annals of Mathematical Statistics* **41**(1), 164–171.

Cappé, O., Moulines, E. and Rydén, T. (2005), *Inference in Hidden Markov Models*, New York: Springer-Verlag.

Casella, G. and George, E. (1992), 'Explaining the Gibbs Sampler', *The American Statistician* **46**(3), 167–174.

Celeux, G., Frühwirth-Schnatter, S. and Robert, C. P. (2018), 'Model Selection for Mixture Models - Perspectives and Strategies', *arXiv preprint arXiv:1812.09885* .

Feller, W. (1943), 'On a General Class of "Contagious" Distributions', *The Annals of Mathematical Statistics* **14**(4), 389–400.

Feller, W. (1948), 'On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions', *The Annals of Mathematical Statistics* **19**(2), 177–189.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, New York: Springer-Verlag.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubbin, D. (2014), *Bayesian Data Analysis*, third edn, Taylor & Francis Group.

Geman, S. and Geman, D. (1984), 'Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.

Grimmett, G. and Stirzaker, D. (2001), *Probability and Random Processes*, third edn, Oxford University Press.

Hamilton, J. D. (1989), 'A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle', *Econometrica* **57**(2), 357–384.

Held, L. and Bové, D. (2014), *Applied Statistical Inference: Likelihood and Bayes*, Springer Science+Business Media.

Huang, C., Lin, Y., Ying, K. and Liang Ku, C. (2011), 'The Solder Paste Printing Process: Critical Parameters, Defect Scenarios, Specifications, and Cost Reduction', *Soldering & Surface Mount Technology* **23**(4), 211–223.

Kay, R. W., Cummins, G., Krebs, T., Lathrop, R., Abraham, E. and Desmulliez, M. (2014), ‘Statistical Analysis of Stencil Technology for Wafer-Level Bumping’, *Soldering & Surface Mount Technology* **26**(2), 71–78.

Lin, J. (2016), ‘On The Dirichlet Distribution’, *Master’s thesis, Queen’s University, Kingston, Ontario, Canada* .

Loy, A., Follett, L. and Hofmann, H. (2016), ‘Variations of Q–Q Plots: The Power of Our Eyes!’, *The American Statistician* **70**(2), 202–214.

Makhoul, J., Starner, T., Schwartz, R. and Chou, G. (1994), ‘On-Line Cursive Handwriting Recognition Using Hidden Markov Models and Statistical Grammars’, *Human Language Technology Conference, Proceedings of the workshop on Human Language Technology* .

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953), ‘Equation of State Calculations by Fast Computing Machines’, *The Journal of Chemical Physics* **21**(6), 1087–1092.

Mirzai, P. and Mårtensson, G. (2019), ‘Robust Reliability Testing for Drop-On-Demand Jet Printing’, *IPC APEX EXPO conference, San Diego, CA* .

Petris, G., Petrone, S. and Campagnoli, P. (2009), *Dynamic Linear Models with R*, New York: Springer-Verlag.

Rabiner, L. (1989), ‘A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition’, *Proceedings of the IEEE* **77**(2), 257–286.

Raftery, A. E. and Lewis, S. (1992), How Many Iterations in the Gibbs Sampler?, *in* ‘In Bayesian Statistics 4’, Oxford University Press, pp. 763–773.

Ross, S. (2010), *Introduction to Probability Models*, 10th edn, Elsevier.

Rydén, T. (2008), ‘EM versus Markov chain Monte Carlo for estimation of hidden Markov models: a computational perspective’, *Bayesian Analysis* **3**(4), 659–688.

Schwarz, G. (1978), ‘Estimating the Dimension of a Model’, *The Annals of Statistics* **6**(2), 461–464.

Sudderth, E. B. (2006), ‘Graphical models for visual object recognition and tracking’, *PhD thesis, MIT* .

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006), ‘Hierarchical Dirichlet Processes’, *Journal of the American Statistical Association* **101**(476), 1566–1581.

Wilk, M. B. and Gnanadesikan, R. (1968), ‘Probability Plotting Methods for the Analysis of Data’, *Biometrika* **55**(1), 1–17.

Zucchini, W. and MacDonald, I. (2009), *Hidden Markov Models for Time Series: An Introduction Using R*, Chapman & Hall/CRC.

A Mathematical Derivations

A.1 The Normal-Inverse-Gamma Distribution

The probability density function of the Normal-inverse-gamma distribution with parameters $(\xi, \lambda, \alpha, \beta)$ is given by

$$P(\mu, \sigma^2 | \xi, \lambda, \alpha, \beta) = \frac{\sqrt{\lambda}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{2\beta + \lambda(\mu - \xi)^2}{2\sigma^2}\right\} .$$

A.2 Derivation of the Joint Probability

Recall that we have a sample $\boldsymbol{y}_{1:n} = (y_1, \dots, y_n)$, and $n_j = \#\{1 \leq k \leq n : X_k = j\}$ is the number of visits to state j in the latent state sequence. The joint probability of the random variables in state j, for $j \in \{1, \dots, m\}$, can be written as

$$\begin{aligned}
\prod_{k:X_k=j} P(y_k|\mu_j, \sigma_j^2) &= \prod_{k:X_k=j} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{(y_k - \mu_j)^2}{2\sigma_j^2}\right\} \\
&\propto (\sigma_j^2)^{-n_j/2} \exp\left\{-\frac{1}{2\sigma_j^2} \sum_{k:X_k=j} (y_k - \mu_j)^2\right\} \\
&= (\sigma_j^2)^{-n_j/2} \exp\left\{-\frac{1}{2\sigma_j^2} \left[\Sigma_{k:X_k=j}(y_k - \bar{y}_j)^2 + n_j(\bar{y}_j - \mu_j)^2\right]\right\} \\
&= (\sigma_j^2)^{-n_j/2} \exp\left\{-\frac{1}{2\sigma_j^2} \left[(n_j - 1)s_j^2 + n_j(\bar{y}_j - \mu_j)^2\right]\right\} ,
\end{aligned}$$

where $s_j^2 = \frac{1}{n_j - 1}\Sigma_{k:X_k=j}(y_k - \bar{y}_j)^2$, and $\bar{y}_j = \Sigma_{k:X_k=j} y_k / n_j$.

A.3 Derivation of Joint Conditional Distribution

The joint conditional distribution can be written as

$$\begin{aligned}
P(\mu_j, \sigma_j^2 | \boldsymbol{y}_{1:n}, \boldsymbol{X}_{1:n}, \dots) &\propto \sigma_j^{-1} (\sigma_j^2)^{-(\upsilon_{0j}/2+1)} \exp\left\{-\frac{1}{2\sigma_j^2} \left[\upsilon_{0j}\sigma_{0j}^2 + \kappa_{0j}(\mu_{0j} - \mu_j)^2\right]\right\} \times \\
&\times (\sigma_j^2)^{-n_j/2} \exp\left\{-\frac{1}{2\sigma_j^2} \left[(n_j - 1)s_j^2 + n_j(\bar{y}_j - \mu_j)^2\right]\right\} \\
&\propto \sigma_j^{-1} (\sigma_j^2)^{-((\upsilon_{0j}+n_j)/2+1)} \exp\left\{-\frac{\upsilon_{0j}\sigma_{0j}^2 + (n_j - 1)s_j^2}{2\sigma_j^2}\right\} \times \\
&\times \exp\left\{-\frac{\kappa_{0j}(\mu_{0j} - \mu_j)^2 + n_j(\bar{y}_j - \mu_j)^2}{2\sigma_j^2}\right\} .
\end{aligned}$$

Let us now expand the last term in the expression, which can be written as

$$\exp\left\{-(\kappa_{0j}+n_j)\left[\mu_j^2 - 2\mu_j\frac{\kappa_{0j}\mu_{0j}+n_j\bar{y}_j}{\kappa_{0j}+n_j} + \frac{\kappa_{0j}\mu_{0j}^2+n_j\bar{y}_j^2}{\kappa_{0j}+n_j}\right]\right\} .$$

Making the variable substitution

$$B = \left[\frac{\kappa_{0j}\mu_{0j}+n_j\bar{y}_j}{\kappa_{0j}+n_j}\right] ,$$

the expression can be written as

$$\begin{aligned}
&\exp\left\{-(\kappa_{0j}+n_j)\left[\mu_j^2 - 2\mu_j B + B^2 - B^2 + \frac{\kappa_{0j}\mu_{0j}^2+n_j\bar{y}_j^2}{\kappa_{0j}+n_j}\right]\right\} \\
&= \exp\left\{-(\kappa_{0j}+n_j)\left[(\mu_j - B)^2 - B^2 + \frac{\kappa_{0j}\mu_{0j}^2+n_j\bar{y}_j^2}{\kappa_{0j}+n_j}\right]\right\} .
\end{aligned}$$

Returning from the variable substitution, the expression can be written as

$$\begin{aligned}
&\exp\left\{-(\kappa_{0j}+n_j)\left[\left(\mu_j - \frac{\kappa_{0j}\mu_{0j}+n_j\bar{y}_j}{\kappa_{0j}+n_j}\right)^2 + \frac{\kappa_{0j}\mu_{0j}^2+n_j\bar{y}_j^2}{\kappa_{0j}+n_j} - \left(\frac{\kappa_{0j}\mu_{0j}+n_j\bar{y}_j}{\kappa_{0j}+n_j}\right)^2\right]\right\} \\
&= \exp\left\{-(\kappa_{0j}+n_j)\left[\left(\mu_j - \frac{\kappa_{0j}\mu_{0j}+n_j\bar{y}_j}{\kappa_{0j}+n_j}\right)^2 + \left(\frac{n_j\kappa_{0j}(\bar{y}_j-\mu_{0j})^2}{(n_j+\kappa_{0j})^2}\right)\right]\right\} .
\end{aligned}$$

Hence, the posterior distribution is proportional to

$$\begin{aligned}
P(\mu_j, \sigma_j^2|\boldsymbol{y}_{1:n}, \boldsymbol{X}_{1:n}, \ldots) &\propto \sigma_j^{-1}(\sigma_j^2)^{-((\upsilon_{0j}+n_j)/2+1)} \exp\left\{-\frac{\upsilon_{0j}\sigma_{0j}^2 + (n_j-1)s_j^2}{2\sigma_j^2}\right\} \times \\
&\times \exp\left\{-\frac{\frac{n_j\kappa_{0j}(\bar{y}_j-\mu_{0j})^2}{n_j+\kappa_{0j}} + (\kappa_{0j}+n_j)\left(\mu_j - \frac{\kappa_{0j}\mu_{0j}+n_j\bar{y}_j}{\kappa_{0j}+n_j}\right)^2}{2\sigma_j^2}\right\} \\
&\propto \text{Scaled N-inv-}\chi^2(\tilde{\mu}_j, \tilde{\sigma}_j^2/\tilde{\kappa}_j; \tilde{\upsilon}_j, \tilde{\sigma}_j^2) ,
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\mu}_j &= \frac{\kappa_{0j}}{\kappa_{0j}+n_j}\mu_{0j} + \frac{n_j}{\kappa_{0j}+n_j}\bar{y}_j \\
\tilde{\kappa}_j &= \kappa_{0j} + n_j \\
\tilde{\upsilon}_j &= \upsilon_{0j} + n_j \\
\tilde{\upsilon}_j\tilde{\sigma}_j^2 &= \upsilon_{0j}\sigma_{0j}^2 + (n_j-1)s_j^2 + \frac{\kappa_{0j}n_j}{\kappa_{0j}+n_j}(\bar{y}_j-\mu_{0j})^2 .
\end{aligned}$$

A.4 Backward Recursion Forward Simulation

The backward recursion forward simulation is a method that can be used to sample the latent state sequence, $\boldsymbol{X}_{1:n}$, during the Gibbs sweeps. The probability of sampling a state, $P(X_k = j)$, has to be computed for every $j \in \{1, \ldots, m\}$, where m is the total number of states. After the probabilities have been computed, the state X_k is sampled in the Gibbs sweep.

Moreover, we allow this process to be a non-homogeneous Markov chain since we allow for different transition probabilities between different time periods.

A.4.1 Derivation of the Probability of Sampling the Initial State

Throughout our derivations, we implicitly condition on the parameter vector, $\boldsymbol{\vartheta}$. The probability of sampling $X_1 = j$ can be written as

$$P(X_1 = j | \boldsymbol{y}_{1:n}) \propto P(X_1 = j, \boldsymbol{y}_{1:n}) \,.$$

Rewriting the expression in terms of conditional probabilities gives

$$P(X_1 = j, \boldsymbol{y}_{1:n}) = P(\boldsymbol{y}_{2:n} | X_1 = j, y_1) P(X_1 = j, y_1) \,.$$

The probability of observing $\boldsymbol{y}_{2:n}$ given X_1 and y_1 only depends on X_1, in accordance with the structure of HMMs. Thus the expression can be written as

$$P(X_1 = j, \boldsymbol{y}_{1:n}) = P(\boldsymbol{y}_{2:n} | X_1 = j) P(X_1 = j, y_1) \,.$$

Finally, using conditional probabilities, the expression can be written as

$$\begin{aligned} P(X_1 = j, \boldsymbol{y}_{1:n}) &= P(\boldsymbol{y}_{2:n} | X_1 = j) P(y_1 | X_1 = j) P(X_1 = j) \\ &= \beta_1(j) f_j(y_1) \pi_j \,. \end{aligned}$$

Since the probabilities are only up to proportional in j, they need to be normalized in order to obtain the correct probabilities.

A.4.2 Derivation of the Probability of Sampling a State Conditional on the Previously Sampled States

Similar to the previous derivation, we implicitly condition on the parameter vector, $\boldsymbol{\vartheta}$. The probability of sampling $X_k = j$ conditional on the history of the chain, $\boldsymbol{X}_{1:k-1}$, and the data, $\boldsymbol{y}_{1:n}$, can be written as

$$P(X_k = j | \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:n}) \propto P(X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:n}) \,.$$

Furthermore, the expression can be factored as

$$\begin{aligned} P(X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:n}) &= P(\boldsymbol{y}_{k+1:n} | X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k}) P(X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k}) \\ &= P(\boldsymbol{y}_{k+1:n} | X_k = j) P(X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k}) \,, \end{aligned}$$

where the last relation holds due to the property of HMMs. Writing the expression as a product of conditional probabilities gives

$$
\begin{aligned}
P(X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:n}) &= P(\boldsymbol{y}_{k+1:n}|X_j = j)P(y_k|X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k-1})\times \\
&\quad \times P(X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k-1}) \\
&= P(\boldsymbol{y}_{k+1:n}|X_j = j)P(y_k|X_k = j)P(X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k-1})\,,
\end{aligned}
$$

where the property holds because y_k only depends on X_k. Using the Markov property, the expression can be written as

$$
\begin{aligned}
P(X_k = j, \boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:n}) &= P(\boldsymbol{y}_{k+1:n}|X_j = j)P(y_k|X_k = j)P(X_k = j|\boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k-1})\times \\
&\quad \times P(\boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k-1}) \\
&= P(\boldsymbol{y}_{k+1:n}|X_j = j)P(y_k|X_k = j)P(X_k = j|X_{k-1} = i)\times \\
&\quad \times P(\boldsymbol{X}_{1:k-1}, \boldsymbol{y}_{1:k-1}) \\
&\propto P(\boldsymbol{y}_{k+1:n}|X_j = j)P(y_k|X_k = j)P(X_k = j|X_{k-1} = i) \\
&= \beta_k(j)f_j(y_k)\gamma_{ij}\,,
\end{aligned}
$$

Similar to the previous derivation, given that the relations are only up to proportional in j, they need to be normalized in order to obtain the correct probabilities.

B Computational Details

The Gibbs sampler is implemented in *Julia*, a high-performance computing language; https://julialang.org/.

JuliaPro-Juno 1.1.0.1. was used as a project interpreter during the implementation of the code. The implemented code can be found on https://github.com/mirzaipatrik/Bayesian_HMM.