Inside-Out: Characterisation of Computed Tomography Noise in Projection and Image Space with Applications to 3D Printing

OxWaSP 2015-16 University of Warwick Cohort

Mini-Project 1 Report

Sherman Ip

Supervisor: Dr. J. Brettschneider (Statistics, Warwick)

Supervisor: Prof. T. Nichols (Statistics, Warwick)

10th June 2016
Abstract

X-ray computed tomography can be used to do quality control on 3D printed samples. However there are sources of error in the 3D printing, how the photons behave and in the X-ray detector. This project aims to find a relationship between the sample mean and sample variance grey values in images obtained from the X-ray detector by fitting linear regressions. In addition, latent variable models such as principle component analysis, factor analysis and the compound Poisson were attempted to be fitted to find sources of variance.
Acknowledgements

• Supervisors: Julia Brettschneider and Tom Nichols

• Inside Out Team: Wilfrid Kendall, Audrey Kueh, Jay Warnett and Clair Barnes


• EPSRC Funding: EP/L016710/1
Contents

1 Introduction 3

2 Literature Review 4
   2.1 X-Ray Production ................................. 5
   2.2 Photon Interactions ............................... 6
   2.3 Detection and Reconstruction ...................... 8

3 About the Data 9
   3.1 Histogram ........................................ 10
   3.2 Normality ......................................... 12
   3.3 Autocorrelation .................................. 12

4 Mean and Variance Relationship 15
   4.1 Weighted Least Squares ............................ 15
      4.1.1 Theory ....................................... 15
      4.1.2 Methods ..................................... 19
      4.1.3 Results ...................................... 19
   4.2 Resampling Weighted Least Squares ............... 21
      4.2.1 Methods ..................................... 21
      4.2.2 Results ...................................... 21
   4.3 Mixture of Linear Regression ...................... 22
      4.3.1 Theory ....................................... 24
      4.3.2 Methods ..................................... 25
      4.3.3 Results ...................................... 26
   4.4 Distance Weighted Least Squares ................. 26
      4.4.1 Methods ..................................... 26
      4.4.2 Results ...................................... 26
Chapter 1

Introduction

3D printing, formally additive manufacturing [1], has recently become state of the art technology which manufactures objects of complicated shapes with very high precision. It was developed in the 1980’s [2] but recently it has been commercialised and is used in medical science [3] and engineering [1]. Because of such a wide range of applications, there has been a need for quality control and this can be done by scanning the 3D printed sample using X-ray computed tomography.

The Inside-Out group at the University of Warwick aims to develop statistical methodology for using X-ray computed tomography to do quality control on 3D printing. Recent research include how the penumbra can be modelled [4], simulating and detecting defects and on the behaviour of dead pixels [5].

Images of X-ray scans of a stationary 3D printed sample were obtained for this project. The aim of this specific project was to find a mean and variance relationship for the grey values. This was done by fitting linear regressions. Separately, latent variable models were fitted to the images to find sources of variance. Latent variable models included principle component analysis, factor analysis and compound Poisson.

This report was written in a Master dissertation style and assumes knowledge up to graduate level in physics and statistics. The mathematics and statistics used can be reviewed in undergraduate textbooks [6] [7], however complicated matrix calculus results will be stated without proof in Appendix B.1.
Computed tomography (CT) scanning is a 3D imaging technique. It does this by reconstructing the geometry of the sample through a series of 2D X-ray images of the sample. The sample rotates after each image taken [8].

Figure 2.1 shows a diagram on how CT scanning works. A 2D image is taken by projecting X-ray photons onto the stationary sample. The photons are then scattered or absorbed by the sample. Some of these photons are then detected by an X-ray detector on the other side of the sample, which produces an image. After an image has been taken, the object rotates and another image is taken. Finally after a number of images, a 3D reconstruction of the object can be estimated [8].

CT scanning was invented by G. Hounsfield [9] in the 1980’s and it was mainly used for medical imaging. The setup for CT scanning is different when scanning patients because the detector and X-ray source rotates around the patient [8]. Recently CT has been used industrially for non-destructive testing in manufacturing [8]. One possible application would be inspecting 3D printed samples.

There are many sources of error in CT scanning [8] and this can cause problems when reconstructing the geometry of the sample. Sources of error include: defects in the detector, environmental noise and the behaviour of photons.

This chapter aims to give brief description on how CT scanning works.
2.1 X-Ray Production

Photons in CT scanning are produced in an X-ray tube. In an X-ray tube, a cathode, consisting of a heated filament, fires projectile electrons through an electric potential to a target which forms the anode [10], as shown in Figure 2.2. Most of the kinetic energy of the projectile electrons is converted into heat however some is converted into electromagnetic radiation. This depends on how the projectile electrons interact with the atoms in the anode [8].

Bremsstrahlung radiation is the result of projectile electrons deaccelerating due to the electrostatic field produced by nucleus of the target. The kinetic energy of the projectile electrons is then converted to electromagnetic radiation to produce X-ray radiation. As a result, the photon energies in bremsstrahlung radiation is a continuous spectrum and can range up to the maximum kinetic energy of the projectile electrons [10].

Characteristic radiation is due to projectile electrons colliding with electrons in the target atom and ionizing them. This produce vacancies in the electron shell and emits photons when the electrons in the target atom drops down back to the ground state. The energy of the emitted radiation is monoenergetic and depends on the binding energy of the target’s atoms [10].

A typical energy spectrum of photons emitted from an X-ray tube is as...
Figure 2.2: An X-ray tube produces photons by firing projectile electrons from a cathode to an anode. Source: G. Michael (2001) [10]. The energy spectrum consist of both bremsstrahlung and characteristic radiation [10].

The voltage and current can be varied in the X-ray tube to produce different energy spectrums and rate of photon production. This can vary the results produced when collecting CT data [8]. Another important factor is the focal spot size because smaller spot sizes produce sharper edges. Larger spot sizes produce unsharp results and this is know as the penumbra effect, as shown in Figure 2.4. However spot sizes too small can produce concentrated heat [11] and can damage the X-ray tube.

2.2 Photon Interactions

Photons emitted by the X-ray tube are projected onto the sample and interacts with it in a number of ways [8].

The sample can effectively absorb photons via the photoelectric effect or pair production [8]. In the photoelectric effect, the photons transfers all its energy to a bounded electron and ejects it from the sample’s atom [12]. In pair production, the photons convert into electron-position pairs by interacting with the Coulomb field of the sample’s atomic nucleus [13]. In
Figure 2.3: A typical energy spectrum of photons emitted from an X-ray tube. The continuous spectrum is the result of Bremmsstrahlung radiation. The peaks are the result of characteristic radiation. Source: G. Michael (2001) [10]

Figure 2.4: Larger focal spot sizes produces unsharp results. This is known as the penumbra effect. Source: F. Welkenhuyzen et al. (2009) [11]
addition to getting absorbed, photons can be scattered by the sample. This happens when photons collide inelastically with and transfers its energy to the sample’s electrons. This process is known as Compton scattering [14].

Suppose a mono-energetic X-ray pencil beam attenuating through an object with varying attenuation coefficient in position $\mu = \mu(x)$. The beam starts at $x = 0$ and is detected at $x = L$, then the attenuation (decrease in X-ray intensity from $I_0$ to $I_1$) is given as [8]

$$I_1 = I_0 \exp \left[ \int_0^L -\mu(x)dx \right]. \tag{2.1}$$

By comparing the intensity before and after attenuation, the integral of the attenuation coefficient along the path of the X-ray can be calculated. Because of the discrete nature of pixels, the integral is usually replaced by a sum [10]. However it was shown that the attenuation coefficient does depend on the energy of the photons [15]. Thus $\mu = \mu(x, E)$ should be made dependent on the energy of the photons [8]. In general low energy photons are more likely to be absorbed than high energy photons, this increases the average energy of the attenuated photons and can be a source of error in CT scanning. This is referred to beam hardening and can cause inaccuracies in Equation (2.1) [10]. This can be reduced by placing a thin sheet of filter to absorb low energy photons [11] or by correcting it in the data analysis stage [10].

### 2.3 Detection and Reconstruction

Most X-ray detectors are scintillator-photodiode detectors. The photons interact with the scintillator material and produce visible light. The visible light is then detected by photodiodes and converts it into electrical current [10]. The detectors used in CT scanning are flat bed scanner which consist of an array of panels of photodiodes [8].

The scale of the 2D images can be calibrated with the use of reference standards [16]. The methods used to reconstruct the 3D geometry of the sample is called the filtered back-projection [10]. The mathematics can be reviewed in [17] and recent methods consider polyenergetic photons [15].
Chapter 3

About the Data

A dataset was given for this project. 100 images of a stationary 3D printed cuboid, with its vertices in the middle, were taken using a Perkin Elmer XRD 1621 digital X-ray detector. These greyscale images were 1996×1996 pixels in size and have a colour depth of 16 bits. This means each pixel can have grey values which take integer values from 0 to $2^{16} - 1$ inclusive. In addition, these images were calibrated so that the air has grey value of 60000. As a result, the units of the grey values have arbitrary units. Unfortunately a scale, in terms of distance, was not recorded. A sample of the images can be viewed in Figure 3.1.

![Image of images](image.png)

Figure 3.1: The 1st, 50th and 99th image in the dataset of X-ray images of a 3D printed cuboid.

The X-ray tube was stated to have the following properties, with the number of significant figures stated as given and no error bars given:
• Voltage = $V = 100$ kV
• Power = $P = 33.0$ W
• Exposure time = $\tau = 500$ ms
• Target material is tungsten.

From this, some additional properties of the X-ray tube was estimated. It is given that the efficiency of photon production $\eta$ is

$$\eta = aVZ = 8.1 \times 10^{-3}$$

(3.1)

where $a = 1.1 \times 10^{-9}$ V$^{-1}$ is a constant and $Z = 74$ is the atomic number of tungsten. The rate of photon production $\nu$ and energy per photon $E$ can be calculated

$$\nu = \frac{\eta P}{Ve} = 1.7 \times 10^{13} \text{ s}^{-1}$$

(3.2)

$$E = Ve = 1.0 \times 10^{2} \text{ keV}$$

(3.3)

where $e$ is the charge per electron.

### 3.1 Histogram

The distribution of all the grey values from all pixels in the dataset is shown in Figure 3.2. It was observed that there were 3 distinct peaks in the histogram at grey values:

• $\sim 2.4 \times 10^{4}$ arb. unit
• $\sim 4.1 \times 10^{4}$ arb. unit
• $\sim 5.0 \times 10^{4}$ arb. unit.

By doing a threshold, as shown in Figure 3.3, it was clear that the 3 main sources for the peaks in the histogram were the 3D printed sample, the background and the foam holder at the bottom of the image.
Figure 3.2: Histogram of all the grey values from all the pixels in the 100 images from the dataset. The grey values ranged from 18 286 to 60 045.

Figure 3.3: A threshold was selected to highlight pixels with certain grey values. Blue: Pixels with grey values less than 34 765. Green: Pixels grey values more than 40 269.
3.2 Normality

For a given pixel, a Normal distribution was fitted on the 100 grey values (one from each image) and the $\chi^2$ goodness of fit test was conducted. This was done for all $1996^2$ pixels as shown in Figure 3.4. When considering the Bonferoni correction [18] for multiple hypothesis testing, the Normal distribution appeared to be a good fit because only 9 pixels have $p$ values less than 10%. In addition, these significant pixels did not seem to cluster and appeared random in space. When ignoring the Bonferoni correction, there was a suspicious patch of low $p$ values on the bottom right. It should be reasonable to fit models where the grey values are Normally distributed, however one should be dubious if unusual results were obtained from the bottom right of the images.

3.3 Autocorrelation

The sample mean and standard error grey value for each of the 100 images were taken and plotted as a time series as shown in Figure 3.5. There was evidence of dependence between samples by looking at the sample autocorrelation plots as shown in Figure 3.6. Because there was a peak at lag 1 in the sample partial autocorrelation plot, the time series could be modelled using an autoregressive model with one parameter. Despite this, the images were assumed to be i.i.d. in this project.
Figure 3.4: $p$ values from the $\chi^2$ goodness of fit test on fitting the Normal distribution on the grey values for each pixel. In b) circled are pixels with $p$ values less than 10%, corrected for multiple hypothesis testing using the Bonferroni correction.
Figure 3.5: The sample mean and standard error grey value for each of the 100 images.

Figure 3.6: Sample autocorrelation and partial autocorrelation of the mean grey value for each image as a time series.
Chapter 4

Mean and Variance Relationship

The aim of this chapter was to do supervised learning on predicting the variance of a pixel’s grey value given its sample mean. This was done by fitting a linear regression to obtain an explanatory model rather than using out of the box machine learning algorithms. Once a good model was found, a single CT scan can be used to quantify the uncertainty on each pixel’s grey value. This uncertainty then can be carried forward when doing volumetric reconstruction.

Figure 4.1 shows the sample variance and sample mean grey value for each pixel. It can be seen that for sample means of \((4.5 \pm 0.5) \times 10^4\), the range of sample variances increased. In addition, there were a few outliers with extremely large sample variances, one example can be seen in Figure 4.2. To take into consideration outliers, a simple weighted least squares regression was done to give less weight to outliers.

4.1 Weighted Least Squares

4.1.1 Theory

The sample variance is a random variable, thus itself has uncertainty. Weighted least squares was used to take into consideration the uncertainty of the sample variance. This was done by giving low weights to data points with high uncertainty on the grey value sample variance.
<table>
<thead>
<tr>
<th>Sample grey value mean (arb. unit)</th>
<th>2</th>
<th>2.5</th>
<th>3</th>
<th>3.5</th>
<th>4</th>
<th>4.5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample mean (arb. unit) × 10^4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample variance (arb. unit^2)</td>
<td>0.5</td>
<td>1</td>
<td>1.5</td>
<td>2</td>
<td>2.5</td>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>Sample variance (arb. unit^2) × 10^{-3}</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 4.1: Frequency density plot of the sample variance against sample mean grey value for all 1 996^2 pixels.

(a) Sample of 7 968 points

(b) Another sample of 7 968 points

Figure 4.2: Scatter plots of the sample variance-mean grey value pairs for a sample of pixels.
Let $S_i^2$ and $\sigma_i^2$ be the sample variance and true variance, respectively, of the $i$th pixel’s grey value for $i = 1, 2, \ldots, m$ and $m = 1996^2$. The sample variance has a sampling distribution such that \cite[pp. 195-198]{7} 

$$
\frac{(n - 1)S_i^2}{\sigma_i^2} \sim \chi^2_{n-1} \tag{4.1}
$$

where $n = 100$ is the number of samples used in estimating the variance. The variance is then

$$2(n - 1) = \frac{(n - 1)^2}{\sigma_i^4} \text{Var} \left( S_i^2 \right)$$

thus

$$\text{Var} \left( S_i^2 \right) = \frac{2\sigma_i^4}{n-1}. \tag{4.2}$$

A good weight for pixel $i$, $w_i$, would be one which is proportional to the precision, or reciprocal of the variance, of $S_i^2$. $\sigma_i^2$ is generally unknown but can be estimated using $S_i^2$. By doing so, the weights for each sample mean-variance pair was chosen to be

$$w_i \propto \frac{1}{(S_i^2)^2}. \tag{4.3}$$

The weights can be introduced into the least squares problem. Let $Y_1, Y_2, \ldots, Y_m$ and $x_1, x_2, \ldots, x_m$ be the sample variance and sample mean grey value respectively. Let $x_1, x_2, \ldots, x_m$ be the $p$ length feature vectors of the sample mean grey values. For example the feature vectors can contain polynomials such that

$$x_i = \begin{pmatrix} 1 \\ x_i \\ x_i^2 \\ \vdots \\ x_i^{p-1} \end{pmatrix}. \tag{4.4}$$

Let $Y_i$ and $x_i$ have a linear relationship such that

$$Y_i = (x_i)^\text{T} \beta + \epsilon_i \tag{4.5}$$

where $\beta$ is a $p$ length parameter vector and $\epsilon_i \sim N(0, \sigma^2_\epsilon)$ for i.i.d. $i = 1, 2, \ldots, m$. The weighted least squares is an optimisation problem where the
objective $T$ is minimised by varying $\beta$

$$T = \frac{\sum_{i=1}^{m} w_i \left( Y_i - (x_i)^T \beta \right)^2}{\sum_{i=1}^{m} w_i}.$$  (4.6)

This can be expanded

$$T = \frac{\sum_{i=1}^{m} w_i \left( Y_i - 2Y_i \beta^T x_i + \beta^T x_i (x_i)^T \beta \right)}{\sum_{i=1}^{m} w_i}.
$$

Using the properties of the trace

$$T = \frac{\sum_{i=1}^{m} w_i \left( Y_i - 2Y_i \text{Tr} \left( \beta^T x_i \right) + \text{Tr} \left( \beta^T x_i (x_i)^T \beta \right) \right)}{\sum_{i=1}^{m} w_i}$$

the differential with respect to $\beta$, using the results in Appendix B.1, is

$$\nabla_\beta T = -2 \frac{\sum_{i=1}^{m} w_i Y_i x_i + 2 \sum_{i=1}^{m} w_i x_i (x_i)^T \beta}{\sum_{i=1}^{m} w_i}.$$  (4.7)

Setting the differential to 0, the value of $\beta$ which minimises the objective is

$$\hat{\beta} = \left[ \sum_{i=1}^{m} w_i x_i (x_i)^T \right]^{-1} \sum_{i=1}^{m} w_i x_i Y_i.$$  (4.8)

Assuming $Y_1, Y_2, \ldots, Y_m$ are independent and have constant variance $\sigma^2\epsilon$, the covariance of $\hat{\beta}$ is

$$\Sigma_{\hat{\beta}} = \text{Cov} \left[ \hat{\beta} \right] = \sigma^2\epsilon \left[ \sum_{i=1}^{m} w_i x_i (x_i)^T \right]^{-1} \left[ \sum_{i=1}^{m} w_i x_i \right]
\left[ \sum_{i=1}^{m} w_i (x_i)^T \right] \left[ \sum_{i=1}^{m} w_i x_i (x_i)^T \right]^{-1}.$$

$\sigma^2\epsilon$ can be estimated using the weighted mean squared error

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{m} w_i \left[ Y_i - (x_i)^T \hat{\beta} \right]^2}{\sum_{i=1}^{m} w_i}.$$  (4.10)
After estimating $\beta$ and $\sigma^2$ and given a data point $x_{\text{test}}$, the predicted grey value variance is

$$\hat{Y} = (x_{\text{test}})^T \hat{\beta} + \epsilon_{\text{test}}$$

(4.11)

with expectation

$$\mathbb{E} \left[ \hat{Y} \right] = (x_{\text{test}})^T \hat{\beta}$$

(4.12)

and variance

$$\text{Var} \left[ \hat{Y} \right] = (x_{\text{test}})^T \Sigma_{\hat{\beta}} x_{\text{test}} + \sigma^2$$

(4.13)

where $\hat{\sigma}^2$ can be used in the estimation of $\text{Var} \left[ \hat{Y} \right]$.

### 4.1.2 Methods

Polynomial features for the sample mean were constructed up to order $k = 6$. The weighted linear regression was then fitted. Before fitting the linear regression, the weights were normalized to have a maximum of 1. Furthermore, each feature was normalized to have mean 0 and standard deviation 1. Afterwards, a constant of 1 was appended to each feature vector to model the intercept of the linear regression.

In addition, the BIC [19, p. 233] was calculated given as

$$\text{BIC} = m \left( \ln 2\pi + \ln \hat{\sigma}^2 + 1 \right) + (k + 1) \ln m$$

(4.14)

for each polynomial order and repeated for 40 different bootstrapped samples. A bootstrapped sample was obtained by randomly resampling $m$ times without replacement from the original dataset [20].

### 4.1.3 Results

The weighted least squares with polynomial features were fitted successfully as shown in Figure 4.3. It appeared successful because the regression goes through the majority of the mean and variance pairs. It was unusual to note that the prediction variance for the 6th order polynomial blew up, further investigation will be needed to explain such behaviour.

The BIC, as shown in Figure 4.4, favoured high order polynomials. Because the BIC favoured more complicated models, this suggested that fitting polynomial features was not a good model. The BIC aimed to seek more and more polynomials so that the model becomes more similar to the true model.
Figure 4.3: Frequency density of the sample mean-variance grey value pairs. Weighted least squares with different orders of polynomial features were fitted. The (red) dotted line represent the 95% prediction interval.
4.2 Resampling Weighted Least Squares

4.2.1 Methods

From the previous chapter, it was clear there were 3 materials, the 3D printed sample, background and foam, which contributed to the peaks in the histogram of the grey values.

The sample mean-variance pairs were resampled from each of the 3 materials, as shown in Figure 4.5. A 1st and 2nd order weighted linear regression were fitted for each material. In addition, the BIC was calculated when fitting such weighted regression on 40 bootstrapped samples [20], for each material.

4.2.2 Results

It was clear the mean-variance relationship is material dependent, this can be seen by the different gradients of the separate regressions in Figure 4.6. Table 4.1 showed that for the 3D printed sample and the foam, the BIC favoured higher order models. However for the background, when taking
Figure 4.5: Highlighted by a square, the resampled pixels from each material. Left: Sample from the background. Middle: Sample from the 3D printed sample. Right: Sample from the foam.

<table>
<thead>
<tr>
<th></th>
<th>Order 1</th>
<th>Order 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>$(4.7421 \pm 0.0003) \times 10^7$</td>
<td>$(4.7413 \pm 0.0003) \times 10^7$</td>
</tr>
<tr>
<td>Background</td>
<td>$(4.177 \pm 0.001) \times 10^6$</td>
<td>$(4.176 \pm 0.001) \times 10^6$</td>
</tr>
<tr>
<td>Foam</td>
<td>$(1.5418 \pm 0.0003) \times 10^6$</td>
<td>$(1.5403 \pm 0.0003) \times 10^6$</td>
</tr>
</tbody>
</table>

Table 4.1: The mean ± standard deviation of the BIC when fitting weighted least squares with polynomial features on 40 bootstrap samples of the resampled mean-variance pairs for each material.

into consideration the uncertainty, the BIC can favour the 1st order model. It was unusual to see the prediction variance to blow for the 2nd order model in the background and foam resampled data.

4.3 Mixture of Linear Regression

In order to take into consideration the different mean-variance relationship for different materials, a mixture of linear regression was tried to see if it captured the different behaviours.
Figure 4.6: Frequency density of the sample mean-variance grey value pairs for each resampled pixels. Weighted linear regression with polynomial features was fitted for each material. The (red) dotted line represent the 95% prediction interval.
4.3.1 Theory

A mixture of linear regression was used of the form

\[ Y_i \sim \begin{cases} 
N \left( (x_i)^T \beta_1, \sigma_1^2 \right) & \text{if } S_i = 1 \\
N \left( (x_i)^T \beta_2, \sigma_2^2 \right) & \text{if } S_i = 2 \\
\vdots & \\
N \left( (x_i)^T \beta_q, \sigma_q^2 \right) & \text{if } S_i = q 
\end{cases} \]  

(4.15)

where \( S_i \) is a latent variable, \( \mathbb{P}(S = r) = \pi_r \) for \( i = 1, 2, \ldots, m \) and \( r = 1, 2, \ldots, q \).

The likelihood is

\[ L = \prod_{i=1}^{m} \sum_{j=1}^{q} p_{Y|S}(y_i|S_i = j) \pi_j \]

thus the log likelihood is

\[ \ln L = \sum_{i=1}^{m} \ln \left[ \sum_{j=1}^{q} p_{Y|S}(y_i|S_i = j) \pi_j \right] \]

\[ \ln L = \sum_{i=1}^{m} \ln \left[ \frac{\pi_j}{\sqrt{2\pi \sigma_j}} \exp \left( -\frac{1}{2} \left( \frac{y_i - (x_i)^T \beta_j}{\sigma_j} \right)^2 \right) \right] . \]  

(4.16)

Such a log likelihood can be maximised by using the EM algorithm \[21\], by treating \( S_i \) as a random latent variable \[22\] pp. 260-261]. The EM algorithm is an iterative algorithm conducting an E step and M step until some convergence condition are met.

In the E step, the posterior distribution is used to put a prediction on the latent variables \( S_i \) for \( i = 1, 2, \ldots, m \) given estimates of the parameters \( \beta_j, \sigma_j^2 \) and \( \pi_j \) for \( j = 1, 2, \ldots, q \). The posterior distribution is given as

\[ r_{i,j} = \mathbb{P}(S_i = j|Y_i = y_i) = \frac{p_{Y|S}(y_i|S_i = j) \pi_j}{\sum_{j'=1}^{q} p_{Y|S}(y_i|S_i = j') \pi_{j'}} \]
and such a quantity is called the responsibility of the $i$th data belonging to the $j$th model.

In the M step, the estimations of the parameters $\beta_j$, $\sigma_j^2$ and $\pi_j$ for $j = 1, 2, \ldots, q$ are updated, using the maximum log likelihood, given the responsibilities $r_{i,j}$ for $i = 1, 2, \ldots, m$. The regression parameters are updated by

$$
\widehat{\beta}_j = \left[ \sum_{i=1}^{m} r_{i,j} x_i (x_i)^T \right]^{-1} \left[ \sum_{i=1}^{m} r_{i,j} y_i x_i \right] 
$$

(4.17)

and this is weighted least squares where the weights are the responsibilities on model $j$. For the rest of the parameters

$$
\widehat{\sigma}_j^2 = \frac{\sum_{i=1}^{m} r_{i,j} (y_i - (x_i)^T \beta_j)^2}{\sum_{i=1}^{m} r_{i,j}} 
$$

(4.18)

and this is the weighted mean squared error. Finally

$$
\widehat{\pi}_j = \frac{\sum_{i=1}^{m} r_{i,j}}{m} 
$$

(4.19)

and this is the average responsibility. The proofs for the M step can be found in Appendix B.2.

### 4.3.2 Methods

Only a mixture of 1st order linear regression was considered. The EM algorithm was initialised by assigning random responsibilities to all models and data points in the E step. The number of mixture was set to $q = 3$ to see if this model captured the 3 materials. 50 EM steps were used.

To make the EM algorithm run properly without problems, a random sample of 10000 sample mean-variance pairs were used in fitting the mixture model.
4.3.3 Results

It was verified that the EM algorithm always increased the log likelihood at every EM step, as can be seen in Figure 4.7.

In Figure 4.8, the 3 regressions did not capture the 3 different materials in the same way resampling weighted least squares did. This was because the mixture probabilities, $\pi_1, \pi_2, \pi_3$, did not depend on the sample mean grey value and did not represent the fact that certain materials have certain grey values. For example, low grey values were more likely to be from the 3D printed sample.

4.4 Distance Weighted Least Squares

4.4.1 Methods

Weighted least squares was used where the weights depended on some distance from some particular grey value. $k = 3$ independent weighted least squares were fitted where the weights are

$$w_i = \exp \left[ -\frac{1}{2} \left( \frac{x_i - \mu_j}{\sigma} \right)^2 \right], \quad (4.21)$$

$\mu_j$ was found using $k$-means \[22\] p. 446 in the sample-mean grey value space and $\sigma$ is a predefined constant, called the Gaussian width.

4.4.2 Results

The fitted model captured the 3D printed sample and foam well because the regression expressed the shape of the histogram, as shown in Figure 4.9. However for the background, the regression had a negative gradient as a result of being influenced by lower sample mean grey values from the foam and 3D printed sample. This did not capture the background in the same way resampled weighted least squares did.

For higher Gaussian widths, the regression became an ordinary least squares regression because all the data points have approximately equal weights.
Figure 4.7: The log likelihood of the mixture of linear regressions at every EM step for 10 different initial values.

Figure 4.8: Fitting a mixture of 3 linear regressions on a sample of 10000 sample mean-variance data. The (red) dotted line represent the 95% prediction interval.
Figure 4.9: 3 independent Gaussian weighted least squares regressions. The weights depended on the distance, from the sample mean to a k-mean sample mean grey value, and some width $\sigma$. The (red) dotted line represent the 95% prediction interval.
Probabilistic latent models was used to model the grey values for each pixel. These latent variables can be used to describe the underlining structure, and perhaps sources of variance. In this chapter, principal component analysis and factor analysis was conducted on the 100 images. At the end, the compound Poisson model will be presented with a toy example.

### 5.1 Principal Component Analysis

Each pixel has a grey value which was treated as a random variable. By estimating the covariance matrix sources of (orthogonal) variance was estimated. This was done by investigating the eigendecomposition representation of the sample covariance matrix. Such a technique is called principal component analysis [22, pp. 329-330].

#### 5.1.1 Theory

Suppose the \( m \times m \) covariance matrix \( \Sigma \) has eigenvalues \( \lambda_i \) and eigenvectors \( \phi_i \) for \( i = 1, 2, \ldots, m \). Then the eigenvectors and eigenvalues are related by

\[
\Sigma \phi_i = \lambda_i \phi_i .
\]  

(5.1)

A closure can be form by pre-multiplying both sides by \( \phi_j^T \) so that

\[
\phi_j^T \Sigma \phi_i = \lambda_i \phi_j^T \phi_i
\]

\[
(\Sigma^T \phi_j)^T \phi_i = \lambda_i \phi_j^T \phi_i .
\]
But given the covariance matrix is symmetric, \( \Sigma = \Sigma^T \), and \( \lambda_j \) is the corresponding eigenvalue of the eigenvector \( \phi_j \), then

\[
(\lambda_j \phi_j)^T \phi_i = \lambda_j \phi_j^T \phi_i
\]

\[
\lambda_j \phi_j^T \phi_i = \lambda_i \phi_j^T \phi_i .
\]

As long as \( \lambda_i \neq \lambda_j \) for \( i \neq j \), then \( \phi_j^T \phi_i = 0 \). Therefore all eigenvalues in the covariance matrix are orthogonal to each other.

Equation (5.1) can be extended to include all eigenvectors and eigenvalues such that

\[
\Sigma \begin{pmatrix}
\uparrow & \uparrow & \cdots & \uparrow \\
\phi_1 & \phi_2 & \cdots & \phi_m \\
\downarrow & \downarrow & \cdots & \downarrow
\end{pmatrix} = \begin{pmatrix}
\uparrow & \uparrow & \cdots & \uparrow \\
\lambda_1 \phi_1 & \lambda_2 \phi_2 & \cdots & \lambda_m \phi_m \\
\downarrow & \downarrow & \cdots & \downarrow
\end{pmatrix} . \quad (5.2)
\]

Transposing both sides

\[
\begin{pmatrix}
\leftarrow & \phi_1 & \rightarrow \\
\leftarrow & \phi_2 & \rightarrow \\
\vdots \\
\leftarrow & \phi_m & \rightarrow
\end{pmatrix} \Sigma = \begin{pmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_m
\end{pmatrix} \begin{pmatrix}
\leftarrow & \phi_1 & \rightarrow \\
\leftarrow & \phi_2 & \rightarrow \\
\vdots \\
\leftarrow & \phi_m & \rightarrow
\end{pmatrix}
\]

but because all eigenvectors are orthogonal to each other

\[
\Sigma = \begin{pmatrix}
\uparrow & \uparrow & \cdots & \uparrow \\
\phi_1 & \phi_2 & \cdots & \phi_m \\
\downarrow & \downarrow & \cdots & \downarrow
\end{pmatrix} \begin{pmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_m
\end{pmatrix} \begin{pmatrix}
\leftarrow & \phi_1 & \rightarrow \\
\leftarrow & \phi_2 & \rightarrow \\
\vdots \\
\leftarrow & \phi_m & \rightarrow
\end{pmatrix} .
\]

This is a weighted sum of the outer products of the eigenvectors, therefore the eigendecomposition is

\[
\Sigma = \sum_{i=1}^{m} \lambda_i \phi_i \phi_i^T . \quad (5.3)
\]

Because this is a weighted sum, the eigenvectors with the biggest eigenvalue will be most responsible for the covariance matrix. \( \phi_i \) are called the principle components.
5.1.2 Methods

The 100 images were shrunk with averaging to a size of $100 \times 100$ pixels using ImageJ [23]. The sample covariance matrix was calculated using the maximum log likelihood estimator so that the 6 eigenvectors with the biggest eigenvalues were obtained. Using 1600 bootstrapped samples [20], the sampling distribution of the 6 largest eigenvalues were sampled.

5.1.3 Results

Figure 5.1 shows the first 6 of the diagonal of the outer product of the individual principle components, they represent the normalised main sources of variance. It was observed that on the first and second principle components, there were sources of variance on the bottom right and top right of the image. This could imply that the sources of variance in the background were not uniformly distributed in space.

The sampling distribution of the eigenvalues are shown in Figure 5.2. Quite clearly there were 2 significant sources of variance because the eigenvalues of the first 2 principle components were much larger than zero when considering its interquartile range.

5.2 Factor Analysis

In factor analysis, grey values for each pixel were modelled as Normally distributed using Normally distributed latent variables. Sources of variance from this model were then isolated and investigated.

5.2.1 Theory

Let $Y$ be a random vector containing $k$ latent variables, called factors, and

$$Y \sim N(0, 1).$$  \hspace{1cm} (5.4)

Let $e$ be a random $p$ vector modelling the uncorrelated intrinsic noise such that

$$e \sim N(0, \Psi)$$  \hspace{1cm} (5.5)

where $\Psi$ is a diagonal matrix.
Figure 5.1: Normalised variance due to the principle components (1 to 6). The sum of all the values in each figure equal to 1.

Figure 5.2: Eigenvalues of the sample covariance matrix. 1600 bootstrap samples were used.
Given a $p \times k$ loading matrix $\Lambda$, the observable data $\mathbf{X}$ can be modelled through

$$\mathbf{X} = \Lambda \mathbf{Y} + \mathbf{e}$$  \hspace{1cm} (5.6)

thus

$$\mathbf{X} \sim N\left(0, \Lambda \Lambda^T + \Psi\right).$$  \hspace{1cm} (5.7)

As a result there are two sources of variance, factor noise from the $\Lambda \Lambda^T$ component and the intrinsic noise $\Psi$. Such a model is called factor analysis [22, pp. 462-469]. The model can be extended for non-zero mean easily by adding an arbitrary constant. $\mathbf{X}$ was used to represent the grey value of all pixels in an image.

The conditional distribution of $\mathbf{X}$ given $\mathbf{Y}$ can be calculated to be

$$\mathbf{X} | \mathbf{Y} \sim N\left(\Lambda \mathbf{Y}, \Psi\right).$$  \hspace{1cm} (5.8)

Using Bayes’ theorem, the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ can be calculated

$$p_{\mathbf{Y} | \mathbf{X}}(\mathbf{y} | \mathbf{x}) \propto p_{\mathbf{X} | \mathbf{Y}}(\mathbf{x} | \mathbf{y}) p_{\mathbf{Y}}(\mathbf{y})$$

$$p_{\mathbf{Y} | \mathbf{X}}(\mathbf{y} | \mathbf{x}) \propto \exp\left[-\frac{1}{2} (\mathbf{x} - \Lambda \mathbf{y})^T \Psi^{-1} (\mathbf{x} - \Lambda \mathbf{y})\right] \exp\left[-\frac{1}{2} \mathbf{y}^T \mathbf{y}\right]$$

$$p_{\mathbf{Y} | \mathbf{X}}(\mathbf{y} | \mathbf{x}) \propto \exp\left[-\frac{1}{2} (-\mathbf{x}^T \Psi^{-1} \mathbf{y} - \mathbf{y}^T \Lambda^T \Psi^{-1} \Lambda \mathbf{y} + \mathbf{y}^T \mathbf{y})\right]$$

$$p_{\mathbf{Y} | \mathbf{X}}(\mathbf{y} | \mathbf{x}) \propto \exp\left[-\frac{1}{2} \left(\mathbf{y}^T \left(\Lambda^T \Psi^{-1} \Lambda + 1\right) \mathbf{y} - 2 \mathbf{y}^T \Lambda^T \Psi^{-1} \mathbf{x}\right)\right]$$

and comparing it with the probability density function of a general multivariate Normal distribution implies

$$\mathbf{Y} | \mathbf{X} \sim N\left(\left(\Lambda^T \Psi^{-1} \Lambda + 1\right)^{-1} \Lambda^T \Psi^{-1} \mathbf{X}, \left(\Lambda^T \Psi^{-1} \Lambda + 1\right)^{-1}\right).$$  \hspace{1cm} (5.9)

Suppose $\mathbf{X} = \mathbf{x}^{(i)}$ was observed for $i = 1, 2, \ldots, n$, then the likelihood is

$$L = \prod_{i=1}^{n} \frac{1}{\|2\pi (\Lambda \Lambda^T + \Psi)\|^{1/2}} \exp\left[-\frac{1}{2} \left(\mathbf{x}^{(i)}\right)^T \left(\Lambda \Lambda^T + \Psi\right)^{-1} \mathbf{x}^{(i)}\right].$$  \hspace{1cm} (5.10)

The log likelihood is then

$$\ln L = -\frac{n}{2} \ln \|2\pi (\Lambda \Lambda^T + \Psi)\| - \frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{x}^{(i)}\right)^T \left(\Lambda \Lambda^T + \Psi\right)^{-1} \mathbf{x}^{(i)}$$
and by using the trace and its properties

\[
\ln L = -\frac{n}{2} \ln \|2\pi (\Lambda \Lambda^T + \Psi)\| - \frac{1}{2} \sum_{i=1}^{n} \text{Tr} \left[ \left(x^{(i)}\right)^T (\Lambda \Lambda^T + \Psi)^{-1} x^{(i)} \right]
\]

\[
\ln L = -\frac{n}{2} \ln \|2\pi (\Lambda \Lambda^T + \Psi)\| - \frac{1}{2} \sum_{i=1}^{n} \text{Tr} \left[ (\Lambda \Lambda^T + \Psi)^{-1} x^{(i)} \left(x^{(i)}\right)^T \right]
\]

\[
\ln L = -\frac{n}{2} \ln \|2\pi (\Lambda \Lambda^T + \Psi)\| - \frac{1}{2} \text{Tr} \left[ (\Lambda \Lambda^T + \Psi)^{-1} n \sum_{i=1}^{n} x^{(i)} \left(x^{(i)}\right)^T \right].
\]

(5.11)

The loading matrix $\Lambda$ and intrinsic noise $\Psi$ can be estimated by maximising the log likelihood. This was done using the EM algorithm [22, pp. 260-261].

The E step estimates the values of the latent variables, $Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)}$, using the conditional expectation of the latent variables given the observed data $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. The parameters $\Lambda$ and $\Psi$ are assumed to be known in this step and are fixed. By using Equation (5.9), each latent variable are predicted by the conditional expectation

\[
E \left[ Y^{(i)} \mid X^{(i)} = x^{(i)} \right] = y^{(i)} = (\Lambda^T \Psi^{-1} \Lambda + 1)^{-1} \Lambda^T \Psi^{-1} x^{(i)}
\]

(5.12) for $i = 1, 2, \ldots, n$. The conditional covariance matrix of the latent variables are estimated by

\[
\text{Cov} \left[ Y^{(i)} \mid X^{(i)} = x^{(i)} \right] = \Sigma_Y = (\Lambda^T \Psi^{-1} \Lambda + 1)^{-1}.
\]

(5.13)

In the M step, the parameters $\Lambda$ and $\Psi$ are varied to maximise the log likelihood, assuming the latent variables are known and fixed. The objective is to maximise is the condition expectation of the log joint likelihood, that is

\[
H = \sum_{i=1}^{n} E \left[ \ln p_{X,Y} \left( x^{(i)}, Y^{(i)} \right) \mid X^{(i)} = x^{(i)} \right].
\]

(5.14)

It is given the $\Lambda$ and $\Psi$ which maximises $H$ is

\[
\hat{\Lambda} = \left( \sum_{i=1}^{n} x^{(i)} \left(y^{(i)}\right)^T \right) \left( n \Sigma_Y + \sum_{i=1}^{n} y^{(i)} \left(y^{(i)}\right)^T \right)^{-1}
\]

(5.15)
and
\[
\hat{\Psi} = \text{diag}\left[\frac{1}{n} \sum_{i=1}^{n} \left( (x^{(i)} - \Lambda y^{(i)}) (x^{(i)} - \Lambda y^{(i)})^T \right) + \Lambda \Sigma_y \Lambda \right].
\] (5.16)

The proofs for the M step can be found in Appendix B.3.

5.2.2 Methods

The 100 images were shrunk with averaging to a size of 100 × 100 pixels using ImageJ [23]. The data was centred so that the mean of each pixel’s gray value was zero. By doing this, factor analysis focused on modelling the variance of the pixel’s gray values.

For a given \( k \) number of factors, the EM algorithm was initialized by randomly assigning each latent variable to be \( N(0, 1) \) distributed and \( \Sigma_y \) to be the estimated covariance of the latent variables.

The EM algorithm ran many times with different initial values. Because evaluating the log likelihood was expensive, the stopping condition was met when a certain amount of E and M steps were taken. The parameters with the highest log likelihood when such condition were met was recorded. The log likelihood at each step was not recorded.

The BIC [19, p. 233] was used to select the best \( k \) to use. The BIC is given as
\[
\text{BIC} = -2 \ln L + 100^2 (k + 1) \ln n.
\] (5.17)

The data was bootstrapped [20] to assess the uncertainty in the BIC.

5.2.3 Results

\( k = 1, 2, \ldots, 15 \) were investigated. The EM algorithm was repeated 80 times for each \( k \) with a stopping condition of 100 EM steps. The BIC for each \( k \) is as shown in Figure 5.3. The number of factors with the lowest BIC was found to be 8. The 8 factor variances were isolated and as shown in Figure 5.4 together with the intrinsic variance as shown in Figure 5.5. It was interesting to note that in Figure 5.4 there were clusters of high variance in most of the factors. This suggested clusters of high variance can appear in the data, thus the variances of the pixel’s gray values can depend on each other spatially. In Figure 5.5 the shape of the 3D printed sample can be seen represented by the low intrinsic variance in the centre. This was evidence that there may
well be a mean and variance relationship because the 3D printed sample has lower sample mean grey values.

The uncertainty of model selection was assessed by investigating how the BIC and optimal $k$ varied for different bootstrapped samples. The BIC for the 100 bootstrapped samples is shown in Figure 5.6. For a given $k$ and bootstrapped sample, the EM algorithm was initialised once to capture the uncertainty if the EM algorithm converges to a global maxima or not. From the figure, quite clearly there was a lot of variability in the BIC and can cause imprecise decisions on model selection. Using all 100 bootstrapped samples, the mean ± standard deviation optimal $k$ was found to be 11 ± 3 factors. The main source of error was down to the small sample size.

## 5.3 Compound Poisson

Consider photons emitted from an X-ray tube as a Poisson process. There has been Poisson latent models developed to consider such behaviour. For example the number of photons detected can be modelled as $Y \sim \text{Poisson}(\lambda)$
Figure 5.4: The 8 individual factor variances with the highest log likelihood found. The units are arbitrary but comparable with the intrinsic variance.

Figure 5.5: The intrinsic variance with the highest log likelihood found. The units are arbitrary but comparable with the factor variances.
and then the grey value can be modelled as $X|Y \sim N(\alpha Y, \beta)$ \cite{24}. Such a model explains that there is a linear relationship between the mean and variance grey value \cite{24}. However this does not consider that these photons have a board energy spectrum as a result of bremsstrahlung and characteristic radiation. It can be shown that a compound Poisson model performs better \cite{25} because it considers the board energy spectrum of the photons. However fitting the model is very difficult using the EM algorithm and approximations have to be made \cite{26}.

5.3.1 Theory

Suppose an image has $m$ pixels and each pixel has a corresponding grey value $X_i$ for $i = 1, 2, \ldots, m$. Let $Y_i$ be the number of photons detected in pixel $i$ and each photon has energy randomly distributed with mean $\mu_i$ and variance $\sigma_i^2$. The number of photons detected can be modelled as a Poisson process such that

$$Y_i \sim \text{Poisson}(\nu_i \tau)$$  \hspace{1cm} (5.18)
where \( \nu_i \) is the post-attenuation rate and \( \tau \) is the exposure time. The total amount of energy measured \( U_i \) is then the sum of the energies of \( Y_i \) photons. For large \( Y_i \), the central limit theorem can be used so that \( U_i \) is approximately Normally distributed so that

\[
U_i | Y_i \sim N \left( Y_i \mu_i, Y_i \sigma_i^2 \right).
\]  

(5.19)

Assuming the grey value is modelling the intensity, rate of energy per unit area, then \( X_i \) can be modelled using

\[
X_i | U_i \sim N \left( \frac{\alpha}{\tau} U_i, \beta_i^2 \right).
\]  

(5.20)

where \( \alpha \) is some positive constant and \( \beta_i^2 \) is the variance of some additive Normal noise.

The full joint probability density function is

\[
p_{X_i, U_i, Y_i} (x_i, u_i, y_i) = p_{X_i | U_i} (x_i | u_i) p_{U_i | Y_i} (u_i | y_i) P(Y_i = y_i)
\]

for any real \( x_i, u_i \) and \( y_i = 0, 1, 2, ... \). By marginalising out \( U_i \), a one layer latent variable model can be obtained. Thus the EM algorithm [22, pp. 260-261] can be used to estimate the unknown parameters \( \nu_i, \mu_i, \sigma_i^2 \) and \( \beta_i^2 \). The one layer latent variable is given as

\[
Y_i \sim \text{Poisson}(\nu_i \tau)
\]  

(5.22)

\[
X_i | Y_i \sim N \left( \frac{\alpha \mu_i}{\tau} Y_i, \frac{\alpha^2 \sigma_i^2}{\tau^2} Y_i + \beta_i^2 \right).
\]  

(5.23)

The proof of this can be found in Appendix B.4.

The marginal expectation and variance [7, pp. 149-151] can be calculated

\[
\mathbb{E}[X_i] = \mathbb{E}[\mathbb{E}[X_i | Y_i]]
\]

39
\[ \mathbb{E}[X_i] = \mathbb{E}\left[ \frac{\alpha Y_i}{\tau} \right] \]
\[ \mathbb{E}[X_i] = \alpha \mu_i \nu_i \] \hspace{1cm} (5.24)

and

\[ \text{Var}[X_i] = \mathbb{E}[\text{Var}[X_i|Y_i]] + \text{Var}[\mathbb{E}[X_i|Y_i]] \]
\[ \text{Var}[X_i] = \mathbb{E}\left[ \frac{\alpha^2 \sigma_i^2}{\tau^2} Y_i + \beta_i^2 \right] + \text{Var}\left[ \frac{\alpha Y_i}{\tau} \right] \]
\[ \text{Var}[X_i] = \frac{\alpha^2 \sigma_i^2}{\tau^2} \nu_i \tau + \beta_i^2 + \frac{\alpha^2 \mu_i^2}{\tau^2} \nu_i \tau \]
\[ \text{Var}[X_i] = \frac{\alpha^2 \nu_i}{\tau} \left( \mu_i^2 + \sigma_i^2 \right) + \beta_i^2 . \] \hspace{1cm} (5.25)

In this model, there are 3 sources of variance from the \( \mu_i^2, \sigma_i^2 \) and \( \beta_i^2 \) terms.

This model can be made simpler by setting \( \beta_i = 0 \) because by doing so the joint probability density function is in the form of an exponential family with sufficient statistics \( X_i^2/Y_i, X_i \) and \( Y_i \). This is proven in Appendix B.5.

However by doing so the joint probability density function must consider the special case of \( Y_i = 0 \) because such a case will drive the variance of \( X_i|Y_i \) to zero. This can be dealt with by using the Dirac delta function [6, pp. 439-443]

\[ p_{X_i,Y_i}(x_i, y_i) = p_{X_i|Y_i}(x_i|y_i) \mathbb{P}(Y_i = y_i) \]
\[ p_{X_i,Y_i}(x_i, y_i) = \begin{cases} 
  e^{-\nu_i \tau} \delta(x) 
  & \text{for } y = 0 \\
  \frac{e^{-\nu_i \tau} (\nu_i \tau)^{y_i}}{y_i! \sqrt{2 \pi y_i \alpha \sigma_i}} \exp \left[ -\frac{1}{2} \frac{(x_i - y_i \mu_i \alpha/\tau)^2}{\alpha^2 y_i \sigma_i^2 / \tau^2} \right] 
  & \text{for } y = 1, 2, \ldots 
\end{cases} \] \hspace{1cm} (5.26)

The conditional probability mass function can be obtained by

\[ \mathbb{P}(Y_i = y_i|X_i = x_i) = \frac{p_{X_i,Y_i}(x_i, y_i)}{p_{X_i}(x_i)} \]

where \( p_{X_i}(x_i) \) is the normalization constant. For the special cases

\[ \mathbb{P}(Y_i = 0|X_i = 0) \propto \delta(0) \]
\[ \mathbb{P}(Y_i = y_i|X_i = 0) \propto \frac{(\nu_i \tau)^{y_i}}{y_i! \sqrt{y_i}} \exp \left[ -\frac{1}{2} \frac{(x_i - y_i \mu_i \alpha/\tau)^2}{\alpha^2 y_i \sigma_i^2 / \tau^2} \right] \] \hspace{1cm} \text{for } y_i = 1, 2, \ldots .
But because $\delta(0)$ is infinite

$$
P(Y_i = y_i | X_i = 0) = \begin{cases} 
1 & \text{for } y_i = 0 \\
0 & \text{for } y_i = 1, 2, \ldots 
\end{cases} \quad (5.27)
$$

and

$$
\mathbb{E}[Y_i | X_i = 0] = 0. \quad (5.28)
$$

$\mathbb{E}[1/Y_i | X_i = 0]$ cannot be defined in this special case.

Another special case is

$$
P(Y_i = 0 | X_i = x) \propto \delta(x) \quad \text{for } x \neq 0
$$

then

$$
P(Y_i = 0 | X_i = x) = 0 \quad \text{for } x \neq 0 \quad (5.29)
$$

For the general case $y_i = 1, 2, \ldots$ and $x_i \neq 0$

$$
P(Y_i = y_i | X_i = x_i) = \frac{1}{p_{X_i}(x_i)} \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i \tau}}{y_i! \sqrt{2\pi y_i \alpha \sigma_i}} \exp \left[ -\frac{1}{2} \frac{\left( x_i - y_i \mu_i \alpha / \tau \right)^2}{\alpha^2 y_i \sigma_i^2 / \tau^2} \right] \quad (5.30)
$$

where

$$
p_{X_i}(x_i) = \sum_{y_i=0}^{\infty} p_{X_i,Y_i}(x_i, y_i)
$$

$$
p_{X_i}(x_i) = \sum_{y_i=1}^{\infty} \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i \tau}}{y_i! \sqrt{2\pi y_i \alpha \sigma_i}} \exp \left[ -\frac{1}{2} \frac{\left( x_i - y_i \mu_i \alpha / \tau \right)^2}{\alpha^2 y_i \sigma_i^2 / \tau^2} \right]. \quad (5.31)
$$

The expectation of the sufficient statistics are then

$$
\mathbb{E}[Y_i | X_i = x_i] = \frac{1}{p_{X_i}(x_i)} \sum_{y_i=1}^{\infty} \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i \tau}}{y_i! \sqrt{2\pi y_i \alpha \sigma_i}} y_i^{1/2} \exp \left[ -\frac{1}{2} \frac{\left( x_i - y_i \mu_i \alpha / \tau \right)^2}{\alpha^2 y_i \sigma_i^2 / \tau^2} \right] \quad (5.32)
$$

and

$$
\mathbb{E}[1/Y_i | X_i = x_i] = \frac{1}{p_{X_i}(x_i)} \sum_{y_i=1}^{\infty} \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i \tau}}{y_i! \sqrt{2\pi y_i \alpha \sigma_i}} y_i^{-3/2} \exp \left[ -\frac{1}{2} \frac{\left( x_i - y_i \mu_i \alpha / \tau \right)^2}{\alpha^2 y_i \sigma_i^2 / \tau^2} \right]. \quad (5.33)
$$
Unfortunately such a sum cannot be simplified. The fastest approximate way to get a value for these expectations is to truncate the sum. Another way to obtain the conditional expectation is to use the sample mean from samples of the conditional distribution, however this could be slower depending on the sampling algorithm.

Because the conditional probability mass function is known up to a constant, rejection sampling can be used to draw conditional samples. Rejection sampling samples exactly but a proposal distribution is needed and this influences the rejection rate, or in other words the efficiency. One possible proposal is the Poisson distribution and there does exist adaptive methods which changes the proposal on the fly \[27\]. Approximate Bayesian computation (ABC) \[28\] can be used instead because \(Y_i\) and \(X_i | Y_i\) can be simulated. ABC draws approximate samples without a proposal but can suffer from high rejection rates.

Suppose for pixel \(i\), the follow grey values have been observed \(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)}\). For each corresponding observable let \(Y_i^{(1)}, Y_i^{(2)}, \ldots, Y_i^{(n)}\) be the Poisson latent variables. For the E step, \(y_i^{(j)}\) and \(\zeta_i^{(j)}\) are estimated given all the parameters using the conditional expectation for \(j = 1, 2, \ldots, n\). That is \(y_i^{(j)} = \mathbb{E} \left[ Y_i^{(j)} \left| X_i^{(j)} = x_i^{(j)} \right. \right] \) and \(\zeta_i^{(j)} = \mathbb{E} \left[ 1/Y_i^{(j)} \left| X_i^{(j)} = x_i^{(j)} \right. \right] \) respectively.

In the M step the parameters \(\nu_i, \mu_i, \sigma_i^2\) are estimated given the latent variables. This is done by maximising the conditional expectation of the log joint likelihood function, that is

\[ H_i = \sum_{j=1}^{n} \mathbb{E} \left[ \ln p_{X_i, Y_i} \left( x_i^{(j)}, y_i^{(j)} \right) \left| X_i^{(j)} = x_i^{(j)} \right. \right] \]  

(5.34)

It is given that the following M step estimators can be obtained:

\[ \hat{\nu}_i = \frac{\sum_{j=1}^{n} y_i^{(j)}}{n \tau} \]  

(5.35)

\[ \hat{\mu}_i = \frac{\tau \sum_{j=1}^{n} x_i^{(j)}}{\alpha \sum_{j=1}^{n} y_i^{(j)}} \]  

(5.36)

\[ \hat{\sigma}_i^2 = \frac{1}{n} \left[ \frac{\tau^2}{\alpha^2} \sum_{j=1}^{n} \left( x_i^{(j)} \right)^2 \zeta_i^{(j)} - \frac{2 \tau \mu_i}{\alpha} \sum_{j=1}^{n} x_i^{(j)} + \mu_i^2 \sum_{j=1}^{n} y_i^{(j)} \right] \]  

(5.37)
where \((x_i^{(j)})^2 \zeta_i^{(j)} = 0\) if \((x_i^{(j)})^2 = 0\). This is derived in Appendix B.6. It is also given that the Cramér-Rao lower bounds \cite{29} are

\[
\text{Var} [\hat{\nu}_i] = \frac{\nu_i}{\tau n} \quad (5.38)
\]

\[
\text{Var} [\hat{\mu}_i] = \frac{\sigma_i^2}{\nu_i \tau n} \quad (5.39)
\]

and

\[
\text{Var} [\hat{\sigma}_i^2] = \frac{2\sigma_i^4}{n} . \quad (5.40)
\]

This is derived in Appendix B.7.

The constant \(\alpha\) only sets the scale of \(\mu_i\) and \(\sigma_i^2\) but can be calibrated so that it has units of keV. From Chapter 3, the following properties of the X-ray tube were stated or calculated:

- Voltage = \(V = 100\) kV
- Power = \(P = 33.0\) W
- Exposure time = \(\tau = 500\) ms
- Efficiency = \(\eta = 8.1 \times 10^{-3}\)
- Target material is tungsten.

Assume, very crudely, the X-ray spreads out and targets the detector with \(1996^2\) pixels only and nothing else, then the total energy detected in one pixel is

\[U = \frac{\eta P \tau}{1996^2} = 2.1 \times 10^8 \text{ keV} \cdot \text{pixel}^{-1}. \quad (5.41)\]

The images were calibrated so that the background has grey value of 60 000. Assuming the air has attenuation coefficient of zero, then \(\alpha\) can be estimated by

\[\alpha = \frac{60 000 \tau}{U} = 1.4 \times 10^{-4} \text{ s} \cdot \text{keV}^{-1} \cdot \text{pixel}. \quad (5.42)\]

This calculation could be improved if the solid angle of the X-ray beam and the distance between the X-ray tube and X-ray detector were given.
5.3.2 Toy Example

A dataset was simulated, with \( n = 100 \) samples, using the follow parameters with arbitrary units: \( \alpha = 1, \tau = 1, \nu = 5, \mu = 3 \) and \( \sigma^2 = 0.1 \). The EM algorithm was initialized by assigning each latent variable such that \( Y^{(j)} \sim \text{Poisson}(\phi) \) for i.i.d. \( j = 1, 2, \ldots, n \). The sums, to work out the expectations and likelihood, were truncated to include 100 terms.

Figure 5.7 showed that the EM algorithm worked for different initial values, \( \phi = 1, 5, 9 \), because the parameters converged towards the true value and the log likelihood always increased. However for \( \phi = 9 \), it appeared the EM algorithm got stuck in a local maxima and needed at least 200 steps for it to converge to the global maxima. This demonstrated that the performance of the EM algorithm depended heavily on the initial values and how many steps were taken.

The EM algorithm was used on 50 different simulated datasets with initial value \( \phi = 4 \). Figure 5.8 verified that the parameters obtained from the EM algorithm converged within the standard error of the true parameters for this particular toy model.

For high \( \nu \tau \), problems arises as more and more terms were needed in the truncated sum. Too many terms in the truncated sum slowed down the E step a lot.
Figure 5.7: The parameters and log likelihood at each step of EM. The EM algorithm was used to estimate the parameters of a single simulated $n = 100$ dataset. The latent variables were initialized Poisson($\phi$) 50 times. Blue: $\phi = 1$, Red: $\phi = 5$, Green: $\phi = 9$. Dotted line: true value
Figure 5.8: The parameters and log likelihood at each step of EM. The EM algorithm was used to estimate the parameters of 50 different simulated $n = 100$ dataset. The latent variables were initialized Poisson(4). Dotted line: true value ± standard error.
Chapter 6

Conclusion

6.1 Mean and Variance Relationship

The relationship between the sample mean and sample variance grey value was modelled by fitting a weighted linear regression. Less weights were given to bigger sample variances as a simple method for robust regression. It was found that the mean and variance relationship depended on the different materials: the 3D printed sample, the background and the foam holder. To take into account the different mean and variance relationships, a mixture of linear regression was fitted but it failed to capture the different behaviours from the different materials. Weighted linear regression, where the weights depended on the difference between the sample mean grey value and some $k$ mean sample mean grey value, was fitted and only captured 2 out of the 3 material mean variance relationships.

There is further work to do in the mean and variance relationship. Firstly it should be investigated whenever transforming the variance will produce a more linear model. For example the log variance or the standard deviation could be used. Furthermore, more features than just polynomials should be used for the sample mean grey values. Lastly, it was established that the sample variance has a scaled $\chi^2$ distribution. Because the $\chi^2$ distribution is in the exponential family, generalised linear models [30] should be used to open up more possibilities to model the relationship between the mean and variance.
6.2 Latent Variable Model

In order to be able to find sources of variance, latent variable models were used as a way to develop explanatory models. PCA and factor analysis were used on a shrunken down dataset and found the 3D printed sample does have a lower sample variance. It was also found that the background variance was not uniform and there were patches of high variance. However shrinking the images down could be a problem because this could of affected the variance of the dataset. In order to improve the data analysis, the computational algorithms should be scaled up so that PCA and factor analysis can be done directly on the full size dataset.

Another explanatory model developed under the light of physics was the compound Poisson distribution. While the compound Poisson distribution was used in the medical physics community [25], fitting the distribution was extremely difficult because the conditional and marginal distributions cannot be written down in closed form. A toy example was provided in this report. In order for the fitting to work computationally, approximations which are cheap to run must be made.
Appendix A

Source Code

MATLAB code can be found on the GitHub account https://github.com/shermanip/oxwasp1
Appendix B

Proofs and Identities

Sections B.1, B.2 and B.3 were heavily influenced by the lectures in probabilistic and unsupervised learning at University College London [31].

B.1 Matrix Calculus

Let $A, B, C$ be matrices with elements

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}$$

and similar for $B$ and $C$.

Let $D_A$ be the gradient operator such that

$$D_A = \begin{pmatrix} \frac{\partial}{\partial a_{1,1}} & \frac{\partial}{\partial a_{1,2}} & \cdots & \frac{\partial}{\partial a_{1,n}} \\ \frac{\partial}{\partial a_{2,1}} & \frac{\partial}{\partial a_{2,2}} & \cdots & \frac{\partial}{\partial a_{2,n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial a_{m,1}} & \frac{\partial}{\partial a_{m,2}} & \cdots & \frac{\partial}{\partial a_{m,n}} \end{pmatrix}$$

then it can be shown that [32]

$$D_A \text{Tr} (A^T B) = B \quad (B.1)$$

$$D_A \text{Tr} (A^T B A C) = B A C + B^T A C^T \quad (B.2)$$
Because a vector is a special case of a matrix, these results can be extended to vectors.

### B.2 M Step for Mixture of Linear Regression

The likelihood for a general mixture model is

\[
\ln L = \sum_{i=1}^{m} \ln \left( \sum_{j=1}^{q} p_{Y_i|S_i} (y_i | S_i = j) \pi_j \right) \tag{B.4}
\]

In the M step, the parameters \( \beta_j, \sigma_j^2 \) and \( \pi_j \) for \( j = 1, 2, \ldots, q \) are updated given the posterior distribution of \( S_i \) for \( i = 1, 2, \ldots, m \). Consider some arbitrary parameter \( \theta_j \) which model \( j \) depends on. Then

\[
\frac{\partial \ln L}{\partial \theta_j} = \sum_{i=1}^{m} \frac{\pi_j}{\sum_{j'=1}^{q} p_{Y_i|S_i} (y_i | S_i = j') \pi_{j'}} \left( \frac{\partial p_{Y_i|S_i} (y_i | S_i = j)}{\partial \theta_j} \right)
\]

\[
\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{m} \frac{\pi_j}{\sum_{j'=1}^{q} p_{Y_i|S_i} (y_i | S_i = j') \pi_{j'}} \left( \frac{\partial \ln p_{Y_i|S_i} (y_i | S_i = j)}{\partial \beta_j} \right)
\]

which is just

\[
\frac{\partial \ln L}{\partial \theta_j} = \sum_{i=1}^{m} r_{i,j} \frac{\partial \ln p_{Y_i|S_i} (y_i | S_i = j)}{\partial \theta_j} \tag{B.5}
\]

Using the above result and the log likelihood in Equation (4.16)

\[
\nabla_{\beta_j} \ln L = \sum_{i=1}^{m} r_{i,j} \nabla_{\beta_j} \ln \left[ \frac{1}{\sqrt{2\pi\sigma_j}} \exp \left( -\frac{1}{2} \left( \frac{y_i - (x_i)^T \beta_j}{\sigma_j} \right)^2 \right) \right]
\]

\[
\nabla_{\beta_j} \ln L = \sum_{i=1}^{m} r_{i,j} \nabla_{\beta_j} \left[ -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma_j^2) - \frac{1}{2} \left( \frac{y_i - (x_i)^T \beta_j}{\sigma_j} \right)^2 \right].
\]

Expanding it out, where \( c_1 \) is some constant,

\[
\nabla_{\beta_j} \ln L = \sum_{i=1}^{m} r_{i,j} \nabla_{\beta_j} \left[ -\frac{1}{2\sigma_j^2} \left( -2y_i \beta_j^T x_i + \beta_j^T x_i (x_i)^T \beta_j \right) + c_1 \right]
\]
and using the properties of the trace
\[ \nabla_{\beta_j} \ln L = \sum_{i=1}^{m} r_{i,j} \nabla_{\beta_j} \left[ -\frac{1}{2\sigma_j^2} \left( -2y_i \operatorname{Tr} (\beta_j^T x_i) + \operatorname{Tr} (\beta_j^T x_i (x_i)^T \beta_j) + c_1 \right) \right] \]
together with the results in Appendix \[B.1\]
\[ \nabla_{\beta_j} \ln L = \sum_{i=1}^{m} -\frac{r_{i,j}}{2\sigma_j^2} \left( -2y_i x_i + 2x_i (x_i)^T \beta_j \right) . \quad (B.6) \]
Setting this to zero
\[ \sum_{i=1}^{m} r_{i,j} y_i x_i = \sum_{i=1}^{m} r_{i,j} x_i (x_i)^T \hat{\beta}_j \]
\[ \widehat{\beta}_j = \left[ \sum_{i=1}^{m} r_{i,j} x_i (x_i)^T \right]^{-1} \left[ \sum_{i=1}^{m} r_{i,j} y_i x_i \right] \quad (B.7) \]
and this is weighted least squares where the weights are the responsibilities on model \( j \).

By re-parametrise \( 1/\sigma_j^2 = \tau_j \)
\[ \frac{\partial \ln L}{\partial \tau_j} = \sum_{i=1}^{m} r_{i,j} \frac{\partial}{\partial \tau_j} \left[ -\frac{1}{2} \ln(2\pi) + \frac{1}{2} \ln(\tau_j) - \frac{\tau_j}{2} \left( y_i - (x_i)^T \beta_j \right)^2 \right] \]
\[ \frac{\partial \ln L}{\partial \tau_j} = \sum_{i=1}^{m} r_{i,j} \left[ \frac{1}{2\tau_j} - \frac{1}{2} \left( y_i - (x_i)^T \beta_j \right)^2 \right] . \quad (B.8) \]
Setting this to zero
\[ \hat{\sigma}_j^2 = \frac{\sum_{i=1}^{m} r_{i,j} \left( y_i - (x_i)^T \beta_j \right)^2}{\sum_{i=1}^{m} r_{i,j}} \quad (B.9) \]
and this is the weighted mean squared error.

To obtain the maximum log likelihood estimator for \( \pi_j \), a Lagrange multiplier must be used to put a constraint such that \( \sum_{j=1}^{q} \pi_j = 1 \). The objective is then
\[ T = \ln L + \lambda \left( \sum_{j=1}^{q} \pi_j - 1 \right) \]
52
\[ \frac{\partial T}{\partial \pi_j} = \frac{\partial}{\partial \pi_j} \sum_{i=1}^{m} \ln \left[ \sum_{j'=1}^{q} p_{Y|S} (y_i | S_i = j') \pi_{j'} \right] + \lambda \]

\[ \frac{\partial T}{\partial \pi_j} = \sum_{i=1}^{m} \frac{p_{Y|S} (y_i | S_i = j)}{\sum_{j'=1}^{q} p_{Y|S} (y_i | S_i = j')} \pi_{j'} + \lambda \]

\[ \frac{\partial T}{\partial \pi_j} = \sum_{i=1}^{m} \frac{r_{i,j}}{\pi_j} + \lambda . \] \hspace{1cm} (B.10)

Setting this to zero

\[ \hat{\pi}_j = - \sum_{i=1}^{m} \frac{r_{i,j}}{\lambda} . \] \hspace{1cm} (B.11)

\( \lambda \) can be determined by putting a constraint on the estimators of \( \pi_j \) for \( j = 1, 2, \ldots, q \)

\[ - \sum_{j=1}^{q} \sum_{i=1}^{m} \frac{r_{i,j}}{\lambda} = 1 \]

\[ - \sum_{i=1}^{m} \sum_{j=1}^{q} r_{i,j} = \lambda \]

\[ - \sum_{i=1}^{m} 1 = \lambda \]

\[ - m = \lambda \] \hspace{1cm} (B.12)

therefore

\[ \hat{\pi}_j = \sum_{i=1}^{m} \frac{r_{i,j}}{m} . \] \hspace{1cm} (B.13)

This is the average responsibility.

**B.3 M Step for Factor Analysis**

The objective is to maximise the condition expectation of the log joint likelihood, that is

\[ H = \sum_{i=1}^{n} \mathbb{E} \left[ \ln p_{X,Y} (x^{(i)}, y^{(i)}) \mid X^{(i)} = x^{(i)} \right] . \]
This can be expressed as

\[ H = \sum_{i=1}^{n} \mathbb{E} \left[ \left( \ln p_{X|Y}(x^{(i)}|Y^{(i)}) + \ln p_Y(Y^{(i)}) \right) \mid X^{(i)} = x^{(i)} \right] \]

Because the marginal distribution of the latent variables does not depend on \( \Lambda \) and \( \Psi \), the last term is not of interest. Using Equation (5.8), where \( c_1 \) is a constant,

\[
H = \sum_{i=1}^{n} \mathbb{E} \left[ \ln \left( \frac{1}{\|2\pi \Psi\|^1/2} \exp \left( -\frac{1}{2} (x^{(i)} - \Lambda Y^{(i)})^T \Psi^{-1} (x^{(i)} - \Lambda Y^{(i)}) \right) \right) \mid X^{(i)} = x^{(i)} \right] + c_1
\]

\[
H = \sum_{i=1}^{n} \mathbb{E} \left[ \left( -\frac{1}{2} \ln \|2\pi \Psi\| \right. \\
\left. -\frac{1}{2} (x^{(i)} - \Lambda Y^{(i)})^T \Psi^{-1} (x^{(i)} - \Lambda Y^{(i)}) \right) \mid X^{(i)} = x^{(i)} \right] + c_1
\]

\[
H = -\frac{1}{2} \sum_{i=1}^{n} \mathbb{E} \left[ \ln \|2\pi \Psi\| + (x^{(i)})^T \Psi^{-1} x^{(i)} \\
-2 (x^{(i)})^T \Psi^{-1} \Lambda Y^{(i)} + (Y^{(i)})^T \Lambda^T \Psi^{-1} \Lambda Y^{(i)} \right] \mid X^{(i)} = x^{(i)} \right] + c_1
\]

By defining

\[ \mathbb{E} \left[ Y^{(i)} \mid X^{(i)} = x^{(i)} \right] = y^{(i)} \quad (B.14) \]

and

\[ \text{Cov} \left[ Y^{(i)} \mid X^{(i)} = x^{(i)} \right] = \Sigma_Y, \quad (B.15) \]

which are obtained from the E step. Then when taking the conditional expectation

\[
H = -\frac{1}{2} \sum_{i=1}^{n} \left[ \ln \|2\pi \Psi\| + (x^{(i)})^T \Psi^{-1} x^{(i)} - 2 (x^{(i)})^T \Psi^{-1} \Lambda y^{(i)} \\
+ \text{Tr} \left( \Lambda^T \Psi^{-1} \Lambda \Sigma_Y \right) + (y^{(i)})^T \Lambda^T \Psi^{-1} \Lambda y^{(i)} \right] + c_1
\]
\[ H = -\frac{1}{2} \sum_{i=1}^{n} \left[ \ln \|2\pi \Psi\| + (x^{(i)})^T \Psi^{-1} x^{(i)} - 2 (x^{(i)})^T \Psi^{-1} \Lambda y^{(i)} \right. \\
\left. + \text{Tr} \left( \Lambda^T \Psi^{-1} \Lambda \Sigma_{Y} \right) + \text{Tr} \left( \Lambda^T \Psi^{-1} \Lambda y^{(i)} (y^{(i)})^T \right) \right] + c_1 \]

\[ H = -\frac{1}{2} \sum_{i=1}^{n} \left[ \ln \|2\pi \Psi\| + (x^{(i)})^T \Psi^{-1} x^{(i)} - 2 (x^{(i)})^T \Psi^{-1} \Lambda y^{(i)} \right. \\
\left. + \text{Tr} \left( \Lambda^T \Psi^{-1} \Lambda \left( \Sigma_{Y} + y^{(i)} (y^{(i)})^T \right) \right) \right] + c_1 \]

\[ H = -\frac{n}{2} \ln \|\Psi\| - \frac{1}{2} \sum_{i=1}^{n} (x^{(i)})^T \Psi^{-1} x^{(i)} + \sum_{i=1}^{n} (x^{(i)})^T \Psi^{-1} \Lambda y^{(i)} \\
- \frac{1}{2} \sum_{i=1}^{n} \text{Tr} \left( \Lambda^T \Psi^{-1} \Lambda \left( \Sigma_{Y} + y^{(i)} (y^{(i)})^T \right) \right) + c_2 , \]

where \( c_2 \) is some constant.

Taking the derivative with respect to \( \Lambda \), using the results from Appendix B.1

\[ D_{\Lambda}H = D_{\Lambda} \left[ \sum_{i=1}^{n} (x^{(i)})^T \Psi^{-1} \Lambda y^{(i)} - \frac{1}{2} \sum_{i=1}^{n} \text{Tr} \left( \Lambda^T \Psi^{-1} \Lambda \left( \Sigma_{Y} + y^{(i)} (y^{(i)})^T \right) \right) \right] \]

\[ D_{\Lambda}H = D_{\Lambda} \left[ \sum_{i=1}^{n} \text{Tr} \left( (y^{(i)})^T \Lambda^T \Psi^{-1} x^{(i)} \right) \\
- \frac{1}{2} \sum_{i=1}^{n} \text{Tr} \left( \Lambda^T \Psi^{-1} \Lambda \left( \Sigma_{Y} + y^{(i)} (y^{(i)})^T \right) \right) \right] \]

\[ D_{\Lambda}H = D_{\Lambda} \left[ \sum_{i=1}^{n} \text{Tr} \left( \Lambda^T \Psi^{-1} x^{(i)} (y^{(i)})^T \right) \\
- \frac{1}{2} \sum_{i=1}^{n} \text{Tr} \left( \Lambda^T \Psi^{-1} \Lambda \left( \Sigma_{Y} + y^{(i)} (y^{(i)})^T \right) \right) \right] \]
\[
D_h H = \sum_{i=1}^{n} \psi^{-1} x^{(i)} (y^{(i)})^T - \sum_{i=1}^{n} \psi^{-1} \Lambda \left( \Sigma_y + y^{(i)} (y^{(i)})^T \right).
\]

Setting the derivative to zero
\[
0 = \sum_{i=1}^{n} \psi^{-1} x^{(i)} (y^{(i)})^T - \sum_{i=1}^{n} \psi^{-1} \tilde{\Lambda} \left( \Sigma_y + y^{(i)} (y^{(i)})^T \right).
\]

\[
0 = \sum_{i=1}^{n} x^{(i)} (y^{(i)})^T - \tilde{\Lambda} \left( n \Sigma_y + \sum_{i=1}^{n} y^{(i)} (y^{(i)})^T \right)
\]

\[
\tilde{\Lambda} = \left( \sum_{i=1}^{n} x^{(i)} (y^{(i)})^T \right) \left( n \Sigma_y + \sum_{i=1}^{n} y^{(i)} (y^{(i)})^T \right)^{-1}.
\] (B.16)

The objective can be re-parametrised using \( T = \Psi^{-1} \) so that
\[
H = -\frac{n}{2} \ln \|T^{-1}\| - \frac{1}{2} \sum_{i=1}^{n} (x^{(i)})^T T x^{(i)} + \sum_{i=1}^{n} (x^{(i)})^T T \Lambda y^{(i)}
\]
\[
- \frac{1}{2} \sum_{i=1}^{n} \text{Tr} \left( \Lambda^T T \Lambda \left( \Sigma_y + y^{(i)} (y^{(i)})^T \right) \right) + c_2.
\]

\( \|T^{-1}\| = 1/\|T\| \) and using the properties of the trace
\[
H = \frac{n}{2} \ln \|T\| - \frac{1}{2} \sum_{i=1}^{n} \text{Tr} \left( T x^{(i)} (x^{(i)})^T \right) + \sum_{i=1}^{n} \text{Tr} \left( T \Lambda y^{(i)} (x^{(i)})^T \right)
\]
\[
- \sum_{i=1}^{n} \frac{1}{2} \text{Tr} \left( T \Lambda \left( \Sigma_y + y^{(i)} (y^{(i)})^T \right) \Lambda^T \right) + c_2.
\]

Taking the derivative with respect to \( T \), using the results from Appendix

\[B.1\]
\[
D_T H = \frac{n}{2} T^{-1} - \frac{1}{2} \sum_{i=1}^{n} x^{(i)} (x^{(i)})^T + \sum_{i=1}^{n} \Lambda y^{(i)} (x^{(i)})^T
\]
\[
- \sum_{i=1}^{n} \frac{1}{2} \Lambda \left( \Sigma_y + y^{(i)} (y^{(i)})^T \right) \Lambda^T.
\]

56
Setting the derivative to zero

\[ 0 = \frac{n}{2} \Psi - \frac{1}{2} \sum_{i=1}^{n} x^{(i)} (x^{(i)})^T + \sum_{i=1}^{n} \Lambda y^{(i)} (x^{(i)})^T \]

\[ - \frac{n}{2} \Lambda \Sigma \Lambda - \frac{1}{2} \sum_{i=1}^{n} y^{(i)} (y^{(i)})^T \Lambda^T \]

\[ 0 = \frac{n}{2} \Psi - \frac{1}{2} \sum_{i=1}^{n} \left( (x^{(i)} - \Lambda y^{(i)}) (x^{(i)} - \Lambda y^{(i)})^T \right) - \frac{n}{2} \Lambda \Sigma \Lambda \]

\[ \Psi = \frac{1}{n} \sum_{i=1}^{n} \left( (x^{(i)} - \Lambda y^{(i)}) (x^{(i)} - \Lambda y^{(i)})^T \right) + \Lambda \Sigma \Lambda \]  

(B.17)

Because \( \Psi \) is a diagonal matrix, only the diagonal entries need to be updated.

### B.4 Obtaining a One-Layer Compound Poisson Model

The full joint probability density function is

\[ p_{X_i, U_i, Y_i}(x_i, u_i, y_i) = p_{X_i|U_i}(x_i|u_i)p_{U_i|Y_i}(u_i|y_i) \mathbb{P}(Y_i = y_i) \]

\[ p_{X_i, U_i, Y_i}(x_i, u_i, y_i) = \frac{1}{\sqrt{2\pi} \beta_i} \exp \left[ -\frac{1}{2} \left( \frac{x_i - \alpha u_i/\tau}{\beta_i} \right)^2 \right] \]

\[ -\frac{1}{\sqrt{2\pi} \sqrt{y_i \sigma_i^2}} \exp \left[ -\frac{1}{2} \left( \frac{u_i - y_i \mu_i}{\sqrt{y_i \sigma_i^2}} \right)^2 \right] e^{-\nu_i \tau (\nu_i \tau)^{y_i}} \frac{y_i!}{y_i!} \]  

(B.18)

for any real \( x_i, u_i \) and \( y_i = 0, 1, 2, \ldots \).

By marginalising out \( U_i \), a one layer latent variable model can be obtained and the unknown parameters \( \nu_i, \mu_i, \sigma_i^2, \beta_i^2 \) can be estimated using the EM algorithm. This can be shown by integrating the full joint probability density

57
function with respect to $u_i$

$$p_{X_i, Y_i}(x_i, y_i) = \int_{u_i = -\infty}^{u_i = \infty} \frac{1}{\sqrt{2\pi \beta_i}} \exp \left[ -\frac{1}{2} \left( \frac{x_i - \alpha u_i / \tau}{\beta_i} \right)^2 \right] \int_{u_i = -\infty}^{u_i = \infty} \frac{1}{\sqrt{2\pi \sqrt{y_i \sigma_i^2}}} \exp \left[ -\frac{1}{2} \left( \frac{u_i - y_i \mu_i}{\sqrt{y_i \sigma_i^2}} \right)^2 \right] e^{-\nu_i \tau (\nu_i \tau)^{y_i}/y_i!} du_i$$

$$p_{X_i, Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau)^{y_i}}}{2\pi \beta_i y_i! \sqrt{y_i \sigma_i^2}} \int_{u_i = -\infty}^{u_i = \infty} \exp \left[ -\frac{1}{2} \left( \frac{x_i - \alpha u_i / \tau}{\beta_i} \right)^2 \right]$$

$$-\frac{1}{2} \left( \frac{u_i - y_i \mu_i}{\sqrt{y_i \sigma_i^2}} \right)^2] \, du_i$$

$$p_{X_i, Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau)^{y_i}}}{2\pi \beta_i y_i! \sqrt{y_i \sigma_i^2}} \int_{u_i = -\infty}^{u_i = \infty} \exp \left[ -\frac{1}{2} \left( \frac{x_i^2 - 2\alpha x_i u_i / \tau + \alpha^2 u_i^2 / \tau^2}{2\beta_i^2} \right) \right.$$

$$\left. - \frac{u_i^2 - 2y_i \mu_i u_i + y_i^2 \mu_i^2}{2y_i \sigma_i^2} \right] \, du_i$$

$$p_{X_i, Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau)^{y_i}}}{2\pi \beta_i y_i! \sqrt{y_i \sigma_i^2}} \int_{u_i = -\infty}^{u_i = \infty} \exp \left[ -\frac{1}{2} \left( \frac{x_i^2 y_i \sigma_i^2 - 2\alpha x_i u_i y_i \sigma_i^2 / \tau + \alpha^2 u_i^2 y_i \sigma_i^2 / \tau^2 + u_i^2 \beta_i^2}{2y_i \sigma_i^2 \beta_i^2} \right) \right.$$

$$\left. - 2y_i \mu_i u_i \beta_i^2 + y_i^2 \mu_i^2 \beta_i^2 \right] \, du_i$$

$$p_{X_i, Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau)^{y_i}}}{2\pi \beta_i y_i! \sqrt{y_i \sigma_i^2}} \int_{u_i = -\infty}^{u_i = \infty} \exp \left[ -\frac{1}{2} \left( \frac{u_i^2 (\beta_i^2 + \alpha^2 y_i \sigma_i^2 / \tau^2) - 2u_i (\alpha x_i y_i \sigma_i^2 / \tau + y_i \mu_i \beta_i^2)}{2y_i \sigma_i^2 \beta_i^2} \right) \right.$$

$$\left. + y_i^2 \mu_i^2 \beta_i^2 + x_i^2 y_i \sigma_i^2 \right] \, du_i .$$
The terms inside the exponential is a quadratic in terms of \( u_i \). The following substitutions are made to make completing the square easier

\[
\begin{align*}
    a_1 &= \beta_i^2 + \alpha^2 y_i \sigma_i^2 / \tau^2 \\
    b_1 &= \alpha x_i y_i \sigma_i^2 / \tau + y_i \mu_i \beta_i^2 \\
    c_1 &= y_i \mu_i \beta_i^2 + x_i y_i \sigma_i^2
\end{align*}
\]  

so that

\[
p_{X_i, Y_i} (x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau) u_i}}{2\pi \beta_i y_i! \sqrt{y_i \sigma_i}} \int_{u_i=-\infty}^{u_i=\infty} \exp \left\{ -\frac{u_i^2 a_1 - 2u_i b_1 + c_1}{2y_i \sigma_i^2 \beta_i^2 / a_1} \right\} du_i.
\]

Completing the square is done by

\[
p_{X_i, Y_i} (x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau) y_i}}{2\pi \beta_i y_i! \sqrt{y_i \sigma_i}} \int_{u_i=-\infty}^{u_i=\infty} \exp \left\{ -\frac{(u_i - b_1/a_1)^2 - b_1^2/a_1 + c_1/a_1}{2y_i \sigma_i^2 \beta_i^2 / a_1} \right\} du_i
\]

\[
p_{X_i, Y_i} (x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau) y_i}}{2\pi \beta_i y_i! \sqrt{y_i \sigma_i}} \exp \left\{ -\frac{b_1^2/a_1^2 + c_1/a_1}{2y_i \sigma_i^2 \beta_i^2 / a_1} \right\} \int_{u_i=-\infty}^{u_i=\infty} \exp \left\{ -\frac{(u_i - b_1/a_1)^2}{2y_i \sigma_i^2 \beta_i^2 / a_1} \right\} du_i.
\]

The integral is now a Gaussian integral which can be evaluated

\[
p_{X_i, Y_i} (x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau) y_i}}{y_i!\sqrt{2\pi \sigma_i}} \exp \left\{ -\frac{b_1^2/a_1^2 + c_1/a_1}{2y_i \sigma_i^2 \beta_i^2 / a_1} \right\} \sqrt{\frac{2\pi y_i \sigma_i^2 \beta_i^2}{a_1}}
\]

\[
p_{X_i, Y_i} (x_i, y_i) = \frac{e^{-\nu_i \tau (\nu_i \tau) y_i}}{y_i!\sqrt{2\pi \sigma_i}} \exp \left\{ -\frac{1}{2a_1} \left( \frac{b_1^2 + a_1 c_1}{y_i \sigma_i^2 \beta_i^2} \right) \right\}.
\]
Substituting in Equations (B.19), (B.20) and (B.21)

\[
p_{X_i,Y_i}(x_i,y_i) = \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i}}{y_i! \sqrt{2\pi \sqrt{\sigma_i^2 \beta_i^2}}} \exp \left[ - \frac{1}{2\alpha_1} \left( -(\alpha x_i y_i \sigma_i^2 \beta_i^2 + y_i \mu_i \beta_i^2)^2 + (\beta_i^2 + \alpha^2 y_i \sigma_i^2 \beta_i^2) \left( y_i^2 \mu_i^2 \beta_i^2 + x_i^2 y_i \sigma_i^2 \beta_i^2 \right) \right) \right]
\]

\[
p_{X_i,Y_i}(x_i,y_i) = \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i}}{y_i! \sqrt{2\pi \sqrt{\sigma_i^2 \beta_i^2}}} \exp \left[ - \frac{1}{2\alpha_1} \left( -\alpha^2 x_i^2 y_i^2 \sigma_i^2 \beta_i^2 / \tau^2 - 2\alpha x_i y_i \sigma_i^2 \mu_i \beta_i^2 / \tau - y_i^2 \mu_i^2 \beta_i^4 \right) \right]
\]

\[
p_{X_i,Y_i}(x_i,y_i) = \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i}}{y_i! \sqrt{2\pi \sqrt{\sigma_i^2 \beta_i^2}}} \exp \left[ - \frac{1}{2\alpha_1} \left( -2\alpha x_i y_i \sigma_i^2 \mu_i \beta_i^2 / \tau + x_i^2 y_i \sigma_i^2 \beta_i^2 + \alpha^2 y_i^2 \sigma_i^2 \mu_i^2 \beta_i^2 / \tau^2 \right) \right]
\]

As before, completing the square can be done by making the following substitutions

\begin{align*}
    a_2 &= y_i \sigma_i^2 \beta_i^2 \\
    b_2 &= \alpha y_i^2 \sigma_i^2 \mu_i \beta_i^2 / \tau \\
    c_2 &= \alpha^2 y_i^2 \sigma_i^2 \mu_i^2 \beta_i^2 / \tau^2
\end{align*}

so that

\[
p_{X_i,Y_i}(x_i,y_i) = \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i}}{y_i! \sqrt{2\pi \sqrt{\sigma_i^2 \beta_i^2}}} \exp \left[ - \frac{1}{2\alpha_1} \left( -2b_2 x_i + a_2 x_i^2 + c_2 \right) \right]
\]

\[
p_{X_i,Y_i}(x_i,y_i) = \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i}}{y_i! \sqrt{2\pi \sqrt{\sigma_i^2 \beta_i^2}}} \exp \left[ - \frac{1}{2\alpha_1} \left( x_i^2 - 2b_2 x_i / a_2 + b_2^2 / a_2^2 - b_2^2 / a_2^2 + c_2 / a_2 \right) \right]
\]

60
\[ p_{X_i, Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i}}{y_i! \sqrt{2\pi \sigma_i}} \exp \left[ -\frac{b_2^2}{a_2^2} + \frac{c_2}{a_2} \right] \exp \left[ -\frac{(x_i - b_2/a_2)^2}{2a_1} \right]. \]

Next is to note that \(-b_2^2/a_2^2 + c_2/a_2 = 0\), this can be shown by

\[
-b_2^2/a_2^2 + c_2/a_2 = -\frac{\alpha^2 y_i^4 \sigma_i^4 \mu_i^2 \beta_i^4 / \tau^2}{y_i \sigma_i^2 \beta_i^2} + \frac{\alpha^2 y_i^4 \sigma_i^2 \mu_i^2 \beta_i^2 / \tau^2}{y_i \sigma_i^2 \beta_i^2}.
\]

\[
-b_2^2/a_2^2 + c_2/a_2 = -\frac{\alpha^2 y_i^2 \mu_i^2}{\tau^2} + \frac{\alpha^2 y_i^2 \mu_i^2}{\tau^2} = b_2^2/a_2^2 + c_2/a_2 = 0. \tag{B.25}
\]

In addition, the following equation can be obtained as well

\[
b_2/a_2 = \frac{\alpha y_i \mu_i}{\tau}. \tag{B.26}
\]

As a result by substituting back in Equation (B.19), the following can be obtained

\[
p_{X_i, Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i}}{y_i! \sqrt{2\pi \beta_i^2 + \alpha^2 y_i \sigma_i^2 / \tau^2}} \exp \left[ -\frac{1}{2} \left( \frac{x_i - y_i \mu_i \alpha / \tau}{\sqrt{\beta_i^2 + \alpha^2 y_i \sigma_i^2 / \tau^2}} \right)^2 \right]. \tag{B.27}
\]

This is the product of a Poisson probability mass function and a Normal probability density function. Finally by inspection the one layer latent model is

\[
Y_i \sim \text{Poisson}(\nu_i \tau) \tag{B.28}
\]

\[
X_i | Y_i \sim \mathcal{N} \left( \frac{\alpha \mu_i}{\tau} Y_i, \frac{\alpha^2 \sigma_i^2}{\tau^2} Y_i + \beta_i^2 \right). \tag{B.29}
\]

### B.5 Joint Compound Poisson is in the Exponential Family

For the \(y = 1, 2, \ldots\) case, the joint probability density function can be rewritten as

\[
p_{X_i, Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau}(\nu_i \tau)^{y_i}}{y_i! \sqrt{2\pi y_i \alpha \sigma_i}} \exp \left[ -\frac{1}{2} \left( \frac{x_i - y_i \mu_i \alpha / \tau}{\alpha^2 y_i \sigma_i^2 / \tau^2} \right)^2 \right].
\]
\[ p_{X_i,Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau}}{\sqrt{2\pi \alpha \sigma_i}} \exp \left[ - \frac{1}{2} \left( \frac{(x_i - y_i \mu_i \alpha / \tau)^2}{\alpha^2 \sigma_i^2 / \tau^2} - y_i \ln(\nu_i \tau) - \frac{1}{2} \ln y_i - \ln y_i! \right) \right] \]

\[ p_{X_i,Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau}}{\sqrt{2\pi \alpha \sigma_i}} \exp \left[ - \frac{\tau^2}{2\alpha^2 \sigma_i^2} \left( \frac{x_i^2}{y_i} - 2 \alpha \mu_i x_i y_i / \tau + \alpha^2 \mu_i^2 y_i^2 / \tau^2 \right) y_i \right. \]
\[ \left. - y_i \ln(\nu_i \tau) - \frac{1}{2} \ln y_i - \ln y_i! \right] \]

\[ p_{X_i,Y_i}(x_i, y_i) = \frac{e^{-\nu_i \tau}}{\sqrt{2\pi \alpha \sigma_i}} \exp \left[ - \frac{\tau^2}{2\alpha^2 \sigma_i^2} \left( \frac{x_i^2}{y_i} - 2 \alpha \mu_i x_i / \tau + \alpha^2 \mu_i^2 y_i / \tau^2 \right) \right. \]
\[ \left. - y_i \ln(\nu_i \tau) - \frac{1}{2} \ln y_i - \ln y_i! \right] \]

This shows that the joint probability density function is a form of an exponential family where the sufficient statistics are \( X_i^2 / Y_i, X_i \) and \( Y_i \).

### B.6 M Step for Compound Poisson

In the M step the parameters \( \nu_i, \mu_i, \sigma_i^2 \) are estimated given the latent variables. This is done by maximising the conditional expectation of the log joint likelihood function, that is

\[ H_i = \sum_{j=1}^{n} \mathbb{E} \left[ \ln p_{X_i,Y_i}(x_i^{(j)}, y_i^{(j)}) \bigg| X_i^{(j)} = x_i^{(j)} \right] \]
Substituting in Equation (5.26), where $c_1$ and $c_2$ are constants,

\[
H_i = \sum_{j=1}^{n} \mathbb{E} \left[ -\nu_i \tau + Y_i^{(j)} \ln(\nu_i \tau) - \frac{1}{2} \ln \left( \frac{\alpha^2 \sigma_i^2}{\tau^2} Y_i^{(j)} \right) - \frac{1}{2} \left( \frac{x_i^{(j)} - Y_i^{(j)} \mu_i / \tau}{\alpha^2 Y_i^{(j)} \sigma_i^2 / \tau^2} \right)^2 \mid X_i^{(j)} = x_i^{(j)} \right] + c_1
\]

\[
H_i = -n\nu_i \tau + \ln(\nu_i \tau) \sum_{j=1}^{n} \mathbb{E} \left[ Y_i^{(j)} \mid X_i^{(j)} = x_i^{(j)} \right] - \frac{n}{2} \ln \left( \frac{\alpha^2 \sigma_i^2}{\tau^2} \right) - \frac{1}{2} \sum_{j=1}^{n} \mathbb{E} \left[ \left( \frac{x_i^{(j)} - Y_i^{(j)} \mu_i / \tau}{\alpha^2 Y_i^{(j)} \sigma_i^2 / \tau^2} \right)^2 \mid X_i^{(j)} = x_i^{(j)} \right] + c_2
\]

\[
H_i = -n\nu_i \tau + \ln(\nu_i \tau) \sum_{j=1}^{n} \mathbb{E} \left[ Y_i^{(j)} \mid X_i^{(j)} = x_i^{(j)} \right] - \frac{n}{2} \ln \left( \frac{\alpha^2 \sigma_i^2}{\tau^2} \right) - \frac{1}{2} \sum_{j=1}^{n} \mathbb{E} \left[ \left( \frac{x_i^{(j)} - 2x_i^{(j)} Y_i^{(j)} \mu_i / \tau + (Y_i^{(j)} \mu_i / \tau)^2}{\alpha^2 Y_i^{(j)} \sigma_i^2 / \tau^2} \right)^2 \mid X_i^{(j)} = x_i^{(j)} \right] + c_2
\]

\[
H_i = -n\nu_i \tau + \ln(\nu_i \tau) \sum_{j=1}^{n} \mathbb{E} \left[ Y_i^{(j)} \mid X_i^{(j)} = x_i^{(j)} \right] - \frac{n}{2} \ln \left( \frac{\alpha^2 \sigma_i^2}{\tau^2} \right) - \frac{\tau^2}{2 \alpha^2 \sigma_i^2} \sum_{j=1}^{n} \mathbb{E} \left[ \left( x_i^{(j)} \right)^2 \mid X_i^{(j)} = x_i^{(j)} \right] + \frac{\mu_i^2}{\alpha \sigma_i^2} \sum_{j=1}^{n} x_i^{(j)} - \frac{\mu_i^2}{2 \sigma_i^2} \sum_{j=1}^{n} \mathbb{E} \left[ Y_i^{(j)} \mid X_i^{(j)} = x_i^{(j)} \right]. \quad \text{(B.30)}
\]

Let $y_i^{(j)} = \mathbb{E} \left[ Y_i^{(j)} \mid X_i^{(j)} = x_i^{(j)} \right]$ and $c_i^{(j)} = \mathbb{E} \left[ 1 / Y_i^{(j)} \mid X_i^{(j)} = x_i^{(j)} \right]$, which

63
was obtained from the E step. Then

\[
H_i = -n \nu_i \tau + \ln(\nu_i \tau) \sum_{j=1}^{n} y_{i}^{(j)} - \frac{n}{2} \ln \left( \frac{\alpha^2 \sigma_i^2}{\tau^2} \right) \\
- \frac{\tau^2}{2\alpha^2 \sigma_i^2} \sum_{j=1}^{n} \left( x_{i}^{(j)} \right)^2 \zeta_{i}^{(j)} + \frac{\tau \mu_i}{\alpha \sigma_i^2} \sum_{j=1}^{n} x_{i}^{(j)} - \frac{\mu_i^2}{2\sigma_i^2} \sum_{j=1}^{n} y_{i}^{(j)} . \tag{B.31}
\]

This can be maximised by setting the partial differentials to zero with respect to \( \nu_i, \mu_i \) and \( \sigma_i^2 \).

Taking the partial derivative with respect to \( \nu_i \)

\[
\frac{\partial H_i}{\partial \nu_i} = -n \tau + \frac{\sum_{j=1}^{n} y_{i}^{(j)}}{\nu_i} . \tag{B.32}
\]

Setting this to zero, the estimator for \( \nu_i \) is

\[
\hat{\nu}_i = \frac{\sum_{j=1}^{n} y_{i}^{(j)}}{n \tau} . \tag{B.33}
\]

Taking the partial derivative with respect to \( \mu_i \)

\[
\frac{\partial H_i}{\partial \mu_i} = \frac{\tau}{\alpha \sigma_i^2} \sum_{j=1}^{n} x_{i}^{(j)} - \frac{\mu_i}{\sigma_i^2} \sum_{j=1}^{n} y_{i}^{(j)} . \tag{B.34}
\]

Setting this to zero, the estimator for \( \mu_i \) is

\[
\hat{\mu}_i = \frac{\tau \sum_{j=1}^{n} x_{i}^{(j)}}{\alpha \sum_{j=1}^{n} y_{i}^{(j)}} . \tag{B.35}
\]

Finally, taking the partial derivative with respect to \( \sigma_i^2 \)

\[
\frac{\partial H_i}{\partial \sigma_i^2} = -\frac{n}{2\sigma_i^2} + \frac{\tau^2}{2\alpha^2 \sigma_i^2} \sum_{j=1}^{n} \left( x_{i}^{(j)} \right)^2 \zeta_{i}^{(j)} - \frac{\tau \mu_i}{\alpha \sigma_i^2} \sum_{j=1}^{n} x_{i}^{(j)} + \frac{\mu_i^2}{2\sigma_i^2} \sum_{j=1}^{n} y_{i}^{(j)} . \tag{B.36}
\]

Setting this to zero

\[
\frac{-n \sigma_i^2}{2} + \frac{\tau^2}{2\alpha^2} \sum_{j=1}^{n} \left( x_{i}^{(j)} \right)^2 \zeta_{i}^{(j)} - \frac{\tau \mu_i}{\alpha} \sum_{j=1}^{n} x_{i}^{(j)} + \frac{\mu_i^2}{2} \sum_{j=1}^{n} y_{i}^{(j)} = 0
\]
and the estimator for $\sigma_i^2$ is

$$\hat{\sigma}_i^2 = \frac{1}{n} \left[ \frac{\tau^2}{\alpha^2} \sum_{j=1}^{n} \left( x_i^{(j)} \right)^2 \zeta_i^{(j)} - \frac{2\tau \mu_i}{\alpha} \sum_{j=1}^{n} x_i^{(j)} + \mu_i^2 \sum_{j=1}^{n} y_i^{(j)} \right].$$  \hspace{1cm} (B.37)

Care should be taken because for $y_i^{(j)} = 0$, $\zeta_i^{(j)}$ is undefined. For this special case it is easiest to just set $\zeta_i^{(j)} = 0$ as though that data point was removed.

### B.7 Cramér-Rao Lower Bounds for the Compound Poisson

Suppose the latent variables are known. The log joint likelihood as a function of random variables is

$$\ln L = -n \nu_{i}\tau + \ln(\nu_{i}\tau) \sum_{j=1}^{n} Y_i^{(j)} - \frac{n}{2} \ln \left( \frac{\alpha^2 \sigma_i^2}{\tau^2} \right)$$

$$- \frac{\tau^2}{2\alpha^2 \sigma_i^2} \sum_{j=1}^{n} \left( X_i^{(j)} \right)^2 / Y_i^{(j)} + \frac{\tau \mu_i}{\alpha \sigma_i^2} \sum_{j=1}^{n} X_i^{(j)} - \frac{\mu_i^2}{2 \sigma_i^2} \sum_{j=1}^{n} Y_i^{(j)}. \hspace{1cm} (B.38)$$

The Fisher’s information matrix is

$$I = -E \begin{pmatrix}
\frac{\partial^2 \ln L}{\partial \nu_i^2} & \frac{\partial^2 \ln L}{\partial \nu_i \partial \mu_i} & \frac{\partial^2 \ln L}{\partial \nu_i \partial \sigma_i^2} \\
\frac{\partial^2 \ln L}{\partial \nu_i \partial \mu_i} & \frac{\partial^2 \ln L}{\partial \mu_i^2} & \frac{\partial^2 \ln L}{\partial \mu_i \partial \sigma_i^2} \\
\frac{\partial^2 \ln L}{\partial \nu_i \partial \sigma_i^2} & \frac{\partial^2 \ln L}{\partial \mu_i \partial \sigma_i^2} & \frac{\partial^2 \ln L}{\partial (\sigma_i^2)^2}
\end{pmatrix}.$$  \hspace{1cm} (B.39)
By substituting in the log joint likelihood, it can be shown that

\[
I = \mathbb{E} \left[ \begin{pmatrix}
\frac{\sum_{j=1}^{n} Y_i^{(j)}}{\nu_i^2} & 0 & 0 \\
0 & \frac{\sum_{j=1}^{n} Y_i^{(j)}}{\sigma_i^2} & \frac{\sum_{j=1}^{n} Y_i^{(j)}}{\sigma_i^2} \\
0 & \frac{\tau \sum_{j=1}^{n} X_i^{(j)}}{\alpha \sigma_i^4} & -\frac{\mu_i}{\sigma_i^4} \sum_{j=1}^{n} Y_i^{(j)} \\
0 & 0 & 0 \\
& & & \end{pmatrix} \right]
\]

\[
= \mathbb{E} \left[ \begin{pmatrix}
\frac{\tau \sum_{j=1}^{n} X_i^{(j)}}{\alpha \sigma_i^4} & -\frac{\mu_i}{\sigma_i^4} \sum_{j=1}^{n} Y_i^{(j)} \\
0 & 0 & 0 \\
& & & \end{pmatrix} \right] 
\]

\[
= \mathbb{E} \left[ \begin{pmatrix}
-n \frac{\nu_i^2}{2 \sigma_i^4} + \frac{\tau^2}{\alpha^2 \sigma_i^6} \sum_{j=1}^{n} \left( X_i^{(j)} \right)^2 / Y_i^{(j)} - \frac{2 \tau \mu_i}{\alpha \sigma_i^6} \sum_{j=1}^{n} X_i^{(j)} + \frac{\mu_i^2}{\sigma_i^6} \sum_{j=1}^{n} Y_i^{(j)} \\
0 & 0 & 0 \\
& & & \end{pmatrix} \right].
\]

(B.40)

Using the results that \( \mathbb{E} \left[ Y_i^{(j)} \right] = \nu_i \tau \) and \( \mathbb{E} \left[ X_i^{(j)} \right] = \alpha \mu_i \), the following can be obtained.

\[
\mathbb{E} \left[ \sum_{j=1}^{n} Y_i^{(j)} \right] = n \nu_i \tau \quad \text{(B.41)}
\]

\[
\mathbb{E} \left[ \sum_{j=1}^{n} X_i^{(j)} \right] = n \alpha \mu_i \nu_i. \quad \text{(B.42)}
\]

Also assuming the E step estimators are unbiased, then

\[
n \sigma_i^2 = \mathbb{E} \left[ \frac{\tau^2}{\alpha^2} \sum_{j=1}^{n} \left( X_i^{(j)} \right)^2 / Y_i^{(j)} - \frac{2 \tau \mu_i}{\alpha \sigma_i^6} \sum_{j=1}^{n} X_i^{(j)} + \frac{\mu_i^2}{\sigma_i^6} \sum_{j=1}^{n} Y_i^{(j)} \right]. \quad \text{(B.43)}
\]

Using these expectations, the Fisher’s information matrix can be calculated to be

\[
I = \begin{pmatrix}
\frac{n \tau}{\nu_i} & 0 & 0 \\
0 & \frac{n \nu_i \tau}{\sigma_i^2} & 0 \\
0 & 0 & \frac{n}{2 \sigma_i^4}
\end{pmatrix}.
\]

(B.44)
The Cramér-Rao bound is

$$\text{Var} \left[ \begin{pmatrix} \hat{\nu}_i \\ \hat{\mu}_i \\ \hat{\sigma}_i^2 \end{pmatrix} \right] \geq \text{diag} \left[ I^{-1} \right] \quad \text{(B.45)}$$

$$\text{Var} \left[ \begin{pmatrix} \hat{\nu}_i \\ \hat{\mu}_i \\ \hat{\sigma}_i^2 \end{pmatrix} \right] \geq \begin{pmatrix} \nu_i \\ n\tau \\ \sigma_i^2 \\ \nu_i \tau \\ 2\sigma_i^4 \\ n \end{pmatrix} \quad \text{(B.46)}$$


