Approximation of Posterior Means and Variances of the Digitised Normal Distribution using Continuous Normal Approximation

Robert Ware[1] and Frank Lad[2]

Abstract

All statistical measurements which represent the values of useful unknown quantities have a realm that is both finite and discrete. Thus our uncertainties about any measurement can be represented by discrete probability mass functions. Nonetheless, common statistical practice treats probability distributions as representable by continuous densities or mixture densities.

Many statistical problems involve the analysis of sequences of observations that the researcher regards exchangeably. Often we wish to find a joint probability mass function over $X_1, X_2, \ldots, X_n$, with interim interest in the sequence of updated probability mass functions $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ for $i = 1, 2, \ldots, n-1$.

We investigate how well continuous conjugate theory can approximate real discrete mass functions in various measurement settings. Interest centres on approximating digital Normal mass functions and digital parametric mixtures with continuous Mixture Normal and Normal-Gamma Mixture Normal distributions for such items as $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$.

Digital mass functions are generated by specifying a finite realm of measurements for a quantity of interest, finding a density value of some specified function at each point, and then normalising the densities over the realm to generate mass values. Both a digitised prior mixing mass function and digitised information transfer function are generated and used, via Bayes' Theorem, to compute posterior mass functions. Approximating posterior densities using continuous conjugate theory are evaluated, and the two sets of results compared.

Key Words: Digital mass functions; Sequential updating; Bayes' Theorem; Mixture Normal; Continuous conjugate theory.

1 Introduction

It is well-known that all statistical measurements which represent the values of useful unknown quantities have a finite and discrete realm of possible measurement val-

[1] Research Fellow, School of Population Health, The University of Queensland

[2] Research Associate, Department of Mathematics and Statistics, University of Canterbury

ues. We denote these as $\mathbf{R}(X) = \{x_1, x_2, \ldots, x_K\}$. Thus our uncertainties about any measurements, if expressed via asserted probability distributions, are represented by discrete probability mass functions. For various historical reasons, common statistical practice treats probability distributions as representable by continuous densities or mixture densities such as the Normal-Gamma mixture-Normal distribution.

If the conditional distribution of X, given μ and σ^2, is Normally distributed with mean μ and variance σ^2, and if μ is also distributed Normally with mean μ_0 and variance τ^2, independent of σ^2, then conventional conjugate mixture-Normal theory tells us that, for a single case, the conditional distribution of X given σ^2, derived by

$$f(x \mid \sigma^2) = \int_{-\infty}^{\infty} f(x \mid \mu, \sigma^2) f(\mu \mid \sigma^2) d\mu, \tag{1}$$

is Normal with mean μ_0 and variance $\sigma^2 + \tau^2$. The distribution of μ given $(X = x)$, whose density will be denoted by $f(\mu \mid x, \sigma^2)$, is also Normal. It has mean $x - \frac{\sigma^2}{\sigma^2+\tau^2}(x - \mu_0)$ and variance $\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$. To summarise this situation symbolically,

$$\begin{aligned} \mu \mid \sigma^2 &\sim N\left(\mu_0, \tau^2\right) & (2)\\ X \mid \mu, \sigma^2 &\sim N\left(\mu, \sigma^2\right) & (3)\\ X \mid \sigma^2 &\sim N\left(\mu_0, \sigma^2 + \tau^2\right) & (4)\\ \text{and} \qquad \mu \mid X = x, \sigma^2 &\sim N\left(x - \frac{\sigma^2}{\sigma^2+\tau^2}(x - \mu_0), \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right). & (5) \end{aligned}$$

If we are to consider these equations in a Bayesian framework we say that $f(\mu \mid \sigma^2)$ is the prior density for μ, $f(x \mid \mu, \sigma^2)$ is the information transfer (or likelihood) function when understood as a function of μ for a fixed value of X, $f(x \mid \sigma^2)$ is the predictive density for X and $f(\mu \mid X = x, \sigma^2)$ is the posterior density for μ.

The purpose of this Report is to investigate how well this continuous conjugate theory, and its extension when σ^2 has an inverted Gamma mixture, can approximate real discrete mass functions when these are representable by digital Normal mass functions and digital parametric mixtures in various measurement settings. The development of large computer memories now allows us to resolve these approximation questions quite accurately for any range of scenarios we may choose.

Digital Normal mass functions are generated by specifying a finite realm of measurements for a quantity of interest, evaluating a Normal density at each point in the realm, and then normalising the density values to generate 'digitised' Normal mass values. Both a digitised prior mixing mass function and a digitised information transfer function are generated and used, via Bayes' Theorem, to compute posterior mass functions. Approximating posterior densities using continuous conjugate theory are evaluated, and the two sets of results compared.

Many statistical problems involve the analysis of sequences of observations that the researcher regards exchangeably. In many cases it is of interest to find a joint probability mass function over $X_1, X_2, \ldots, X_n$, with interim interest in the sequence of updated probability mass functions $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ for $i = 1, \ldots, n-1$. The bold symbol $\mathbf{X}_i$ denotes the vector of quantities $(X_1, \ldots, X_i)$. The lower case $\mathbf{x}_i$ is the vector of their realised values, $(x_1, \ldots, x_i)$. In this context we are interested in the approximate adequacy of the continuous conjugate theory for such items as $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$.

This Report begins with an extended description of the programming strategy used to solve a very simplified problem. Once we are clear on the setup and structure of the procedure we can address more involved mixture problems using a notation that will have become concise. In Section 2 we consider our problem when the prior density is fully specified. In Section 3 we consider the same problem when a hierarchical Bayes model is assessed. Each of these two main Sections begins with an investigation into how we can construct useful parametric families of discrete mass functions by 'digitising' well-known continuous distributions in the context of problems that are now conventionally set up as mixture-Normal distributions (Section 2), or Normal-Gamma mixture-Normal distributions (Section 3). We then show how we can use Bayes' Theorem to compute means and variances of posterior mass functions. We review how conventional conjugate theory is used to generate approximate posterior densities. Finally, we apply the theories of the previous subsections to both the mixture-Normal and Normal-Gamma mixture-Normal distributions. Different parameter values are used to compare results obtained from the use of digitised distributions and results gathered through the use of conventional conjugate theory. The final Section contains a summary of the work.

The main achievement of this Report is to formalise a computing strategy that can be applied to many functional forms. Ware and Lad (2003) apply this computational procedure to various families of extreme value distributions, relying on this present Report for a more detailed discussion of construction details. We begin with a complete analysis of the simplest case in all its details.

2 Approximation of Posterior Means and Variances when a One-Stage Prior Distribution is Specified

One way of constructing useful parametric families of discrete mass functions is to ‘digitise’ well-known continuous distributions, such as the Normal, in the context of problems that are now conventionally set up as mixture-Normal distributions as described above. The construction procedure amounts to evaluating a specified Normal density over the finite realm of discrete possible values of the measured quantities, and normalising to make the density values sum to 1. We now consider how this can work.

2.1 Discrete Mass Functions Characterised as Digitised Normal Densities

The simplest problem of the form described above arises when the prior density is fully specified. We assume that the variance parameter of the conditional distribution of observations is known, and is σ^2. We assume that the mean parameter is unknown, and we assess it as being Normally distributed with mean μ_0 and variance τ^2. In this case we need only work with the probability mass functions $f(x \mid \mu, \sigma^2)$ and $f(\mu \mid \sigma^2)$, which we want to characterise as digitised Normal densities. We begin by specifying a finite realm of measurements for a quantity of interest, evaluating the density at each point in that realm and then normalising the density values to generate the probability mass function.

2.1.1 Forming Matrices of Mass Values

To generate the digitised spaces with which we are concerned we start by defining two realms. One will represent the possible measurement values of X, and will be denoted $\mathbf{R}(X)$. The elements of $\mathbf{R}(X)$ are determined by the measuring device used to identify X. The other realm will represent the location of the digitised mixture function with parameter μ, and is denoted $\mathbf{R}(\mu)$. We shall delimit the elements of $\mathbf{R}(\mu)$ only to a degree of fineness that really interests us. We must be careful to ensure that each realm includes every possible value that realistically could be attained by its subject. We shall specify a general computing structure for such spaces that will allow a range of specific situations by varying the distance between the elements within each of the two realms. For the purposes of this Report both $\mathbf{R}(X)$ and $\mathbf{R}(\mu)$ will be confined to being symmetric about 0.

Once we have identified each of these realms we use them to form a matrix of mass values for $f(x \mid \mu, \sigma^2)$. The matrix has rows corresponding to the candidate X values and columns corresponding to the possible values for μ. At each grid-point on our matrix we evaluate $f(x \mid \mu, \sigma^2)$ from the density corresponding to the μ of that column. In this current setup the value of σ^2 is fixed. The columns are normalised to sum to 1, as later they are used in the computation of the posterior mass function $f(\mu \mid X = x, \sigma^2)$. Each column is a digital Normal mass function for X conditioned on the corresponding value of μ.

The second task is to construct a matrix of mass values to represent the prior mixing mass function. To aid our later computations, matrices $f(x \mid \mu, \sigma^2)$ and $f(\mu \mid \sigma^2)$ should be the same size. We want to characterise $f(\mu \mid \sigma^2)$ as a digital $N(\mu_0, \tau^2)$ curve. Note that μ_0 and τ^2 are preselected values. At each element contained in $\mathbf{R}(\mu)$ we evaluate $f(\mu \mid \sigma^2)$. This vector of density values is then normalised to produce a digital mass function. Note that the length of the row vector $f(\mu \mid \sigma^2)$ is equal to the number of columns in matrix $f(x \mid \mu, \sigma^2)$. Finally, row vector $f(\mu \mid \sigma^2)$ is replicated and tiled to produce a matrix which is the same size as $f(x \mid \mu, \sigma^2)$. The vector $f(\mu \mid \sigma^2)$ is not removed from the memory of the computer, as it is required for use further on in our computations.

2.1.2 The Information Transfer Function

An information transfer function (ITF) is used to update our uncertain knowledge about X_{i+1} given we know that $\mathbf{X}_i = \mathbf{x}_i$ has occurred. The common parlance for what we call an ITF is a "likelihood function". We prefer the use of the term "information transfer function" to accentuate the realisation that there really is no "true value of μ" to be estimated, only sequential values of X's to be observed. In the context of opinions expressed by the exchangeable distribution, the ITF is the form in which the observation of one X allows us to infer something about the next — thus the name "information transfer function". (See pp. 397–999 of Lad (1996) for further discussion.)

The ITF can be calculated for each potential observation value. Each row of matrix $f(x \mid \mu, \sigma^2)$ corresponds to the ITF for one candidate X value. Thus for any observed value of $X = x$, we merely read the corresponding row of the $f(x \mid \mu, \sigma^2)$ matrix of mass values as the ITF.

2.1.3 Generating Sequences of Observation Values

We want to generate $\mathbf{X}_N$, a vector of length N whose elements represent observations from a distribution. These observations will be used, via the ITF, to compute such items as $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$.

To simulate an observation vector we randomly select elements from $\mathbf{R}(X)$. For the purpose of generating observations we assume that μ is known to us. The variance parameter, σ^2, is also a predetermined value. To generate a sequence of observation values we follow these steps:

1. Extract the column of $f(x \mid \mu, \sigma^2)$ which corresponds to the chosen μ. This is a mass function characterised as a digitised $N(\mu, \sigma^2)$ density. The number of elements in the mass function is the same as the size of the realm of X.

2. Form a cumulative mass function, $F(x \mid \mu, \sigma^2)$, using the mass function from the preceding step.

3. Generate $\mathbf{U}_N$, a vector of size N consisting of $U(0, 1)$ random variables.

4. Find the smallest element of $\mathbf{R}(X)$ for which $F(x \mid \mu, \sigma^2) > U_i$ is true. This is X_i. Each element of $\mathbf{U}_N$ corresponds to a constituent of the cumulative mass function, and thus to a member of the realm of X.

5. Repeat steps 1–4 until the sequence of observations, $\mathbf{X}_N$, is complete.

To this point we have constructed:

(a) $\mathbf{R}(X)$, a vector to represent the realm measurement values of X,

(b) $\mathbf{R}(\mu)$, a vector to represent the realm of the location of the digitised mixture function with parameter μ,

(c) $f(x \mid \mu, \sigma^2)$, a matrix of mass values whose columns correspond to conditional probability mass functions for X given different values of μ, and whose rows correspond to ITF's for the various observation values of X,

(d) $f(\mu \mid \sigma^2)$, a matrix of mass values with identical rows, each of which designates a digitised Normal mass function. The size of this matrix is equal to that of $f(x \mid \mu, \sigma^2)$,

(e) $\mathbf{X}_N$, an observation vector whose elements have been generated randomly from digitised Normal mass function $f(x \mid \mu, \sigma^2)$, where μ has been assumed known.

In the course of this study we shall perform this routine using different values of μ. Remember that our objective is, via ITF's, to see how closely continuous conjugate theory can approximate real discrete mass functions in various measurement settings. We can do this by comparing the sequential predictive distributions for components of $\mathbf{X}_N$ attained through using conjugate theory and through using real digital mass function computations. We can also compare the posterior means and variances of the posterior digitised Normal mixtures with approximate values that we compute using standard conjugate methods. In other words, how do the means and variances of $X_{i+1} \mid (\mathbf{X}_i = \mathbf{x}_i)$ compare using exact and approximate methods?

2.2 Computing Posterior Means and Variances of Posterior Discrete Densities

To find exact values for $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ for the discrete mass functions characterised as digitised Normal densities we should compute

$$f\left(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2\right) = \sum_{\mu} f\left(x_{i+1} \mid \mu, \mathbf{X}_i = \mathbf{x}_i, \sigma^2\right) f\left(\mu \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2\right). \qquad (6)$$

The first function in this sum of products of functions does not change. It is always represented by the columns of the matrix $f(x \mid \mu, \sigma^2)$ we have just discussed. The second function is the sum of Equations corresponding to (2) and (3) and changes with each observed value of $\mathbf{X}_i = \mathbf{x}_i$ according to Bayes' Theorem. This implies

$$f(\mu \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2) = \frac{f(\mathbf{X}_i = \mathbf{x}_i \mid \mu, \sigma^2) f(\mu \mid \sigma^2)}{\sum_{\mu} f(\mathbf{X}_i = \mathbf{x}_i \mid \mu, \sigma^2) f(\mu \mid \sigma^2)} \qquad \text{for} \quad i = 1, 2, \ldots, N-1. \qquad (7)$$

Bayes' Theorem is a computational formula for determining posterior probability distributions conditional upon observing the data $\mathbf{X}_i = \mathbf{x}_i$. The posterior probability distribution reflects our revised mixing function over values of μ in light of the knowledge that $\mathbf{X}_i = \mathbf{x}_i$ has occurred.

To compute $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ we merely compute the mean and variance of the appropriate $f_{X_{i+1}}(x \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2)$ that has been computed from the observed data. The formulae

$$\begin{aligned} E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) &= \sum_{x} x f_{X_{i+1}}\left(x \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2\right) & (8) \\ \text{and} \qquad V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) &= \sum_{x} x^2 f_{X_{i+1}}\left(x \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2\right) - \\ & \qquad \left[\sum_{x} x f_{X_{i+1}}\left(x \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2\right)\right]^2, & (9) \end{aligned}$$

are implemented by multiplying the elements of the realm of X (in Equation 9 the elements of $\mathbf{R}(X)$ are squared) element-wise with $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2)$ and then summing the products.

It is easy to find $E(X_1)$ and $V(X_1)$. The matrices $f(\mu \mid \sigma^2)$ and $f(x \mid \mu, \sigma^2)$ both reside in the computer via the calculations described in Section 2.1. Use Equation 6 to find $f(x_1 \mid \sigma^2)$ and then apply Equations 8 and 9.

To find items $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$, where $i \geq 1$, we 'observe' X_i and use the information this observation gives us to form the ITF and implement Bayes' Theorem, thus finding the updated mixing function $f(\mu \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2)$. We are now in position to re-apply Equation 6 and compute the updated predictive mass function, and thus find $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ exactly! In particular, to find $E(X_2 \mid X_1 = x_1)$ and $V(X_2 \mid X_1 = x_1)$:

1. Observe $X_1 = x_1$.

2. Extract the row corresponding to X_1 from the matrix $f(x \mid \mu, \sigma^2)$. This is the ITF through μ from $X_1 = x_1$ to X_2.

3. Implement Bayes' Theorem to update the mixing function, $f(\mu \mid X_1 = x_1, \sigma^2)$. This involves multiplying vectors $f(X_1 = x_1 \mid \mu, \sigma^2)$ and $f(\mu \mid \sigma^2)$ element-wise, and normalising.

4. Replicate and tile vector $f(\mu \mid X_1 = x_1, \sigma^2)$ to form a matrix which has the same dimensions as matrix $f(x \mid \mu, \sigma^2)$.

5. Calculate $f(x_2 \mid X_1 = x_1, \sigma^2)$, the updated predictive mass function, according to Equation 6. Matrix $f(x \mid \mu, \sigma^2)$ is multiplied element-wise with the updated mixing function matrix, and the columns are summed. The resulting vector has length equal to the size of $\mathbf{R}(X)$.

6. Compute $E(X_2 \mid X_1 = x_1)$ and $V(X_2 \mid X_1 = x_1)$ using Equations 8 and 9.

Repeat Steps 1–6 as many times as required to obtain $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$.

Note that for the purpose of this Report the term "vector" refers to a one-dimensional array of size $a \times 1 \times 1$ where $a > 1$. The term "matrix" refers to a two-dimensional array of size $a \times b \times 1$ where $a, b > 1$. The term "array" refers to a three-dimensional array of size $a \times b \times c$ where $a, b, c > 1$.

2.3 Conventional Conjugate Mixture-Normal Theory

Conventional conjugate mixture-Normal theory tells us that if $\mu \mid \sigma^2 \sim N(\mu_0, \tau^2)$ and $X \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$, then $X \mid \sigma^2 \sim N(\mu_0, \sigma^2 + \tau^2)$. In fact the joint exchangeable distribution for the entire sequence of X's is multivariate Normal:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_K \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_0 \\ \mu_0 \\ \vdots \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \sigma^2+\tau^2 & \tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \tau^2 & \sigma^2+\tau^2 & & \tau^2 \\ \vdots & \vdots & & \ddots & \\ \tau^2 & \tau^2 & \tau^2 & & \sigma^2+\tau^2 \end{bmatrix}\right). \quad (10)$$

Which can be written in the form $\mathbf{X}_K \sim N_K(\mu_0 \mathbf{1}_K, \sigma^2 \mathbf{I}_K + \tau^2 \mathbf{1}_{K,K})$. For any individual X_i, the mean is equal to μ_0, the variance equals $\sigma^2 + \tau^2$ and the covariance with any other X_j equals τ^2.

Standard multivariate Normal theory says that we assert $\mathbf{X}_K \sim N_K(\mu_{\mathbf{0}}, \mathbf{\Sigma})$, and if the K-dimensional vector $\mathbf{X}$ is partitioned into two sub-vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, where $(\mathbf{X}_1, \mathbf{X}_2)$ is any partition of $\mathbf{X}$ into its first K_1 and remaining K_2 components, then the conditional distribution for $\mathbf{X}_2 \mid (\mathbf{X}_1 = \mathbf{x}_1)$ is

$$\mathbf{X}_2 \mid (\mathbf{X}_1 = \mathbf{x}_1) \sim N_{K_2}\left[\mu_{\mathbf{0}_2} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \mu_{\mathbf{0}_1}), \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\right]. \quad (11)$$

Applying this multivariate Normal result to the exchangeable Normal distribution tells us that the conditional density for $\mathbf{X}_2 \mid (\mathbf{X}_1 = \mathbf{x}_1)$ can be assessed as

$$\mathbf{X}_2 \mid (\mathbf{X}_1 = \mathbf{x}_1) \sim N_{K_2}\left[\frac{\sigma^2\mu_0 + \tau^2 \sum_{i=1}^{K_1} x_{1_i}}{(\sigma^2 + K_1\tau^2)}\mathbf{1}_{K_2}, \sigma^2\mathbf{I}_{K_2} + \frac{\sigma^2\tau^2}{(\sigma^2 + K_1\tau^2)}\mathbf{1}_{K_2,K_2}\right]. \quad (12)$$

For details see the text of Lad (1996, pp. 375–376, 387–388).

In the specific application to our problem involving the forecast of X_{i+1} given $(\mathbf{X}_i = \mathbf{x}_i)$ these general results apply with the partitioned vector $\mathbf{X}_1$ equal to the condition vector $(\mathbf{X}_i = \mathbf{x}_i)$ and the partitioned vector $\mathbf{X}_2$ equal to the quantity X_{i+1}. The two items that we are most interested in are the posterior conditional expectation of X_{i+1} given the observations $(\mathbf{X}_i = \mathbf{x}_i)$, and the posterior conditional variance of X_{i+1} given the observations $(\mathbf{X}_i = \mathbf{x}_i)$.

The conditional expectation reduces to

$$E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) = \left(\sigma^2 + i\tau^2\right)^{-1}\left(\sigma^2\mu_0 + i\tau^2\bar{x}_i\right), \quad (13)$$

where $\bar{x}_i$ is the average of the observed $\mathbf{X}_i$. We observe that as the number of observations increases, the relative weight on the prior mean, and on each individual observation value, decreases. However, the relative weight on the observed mean, $\bar{x}_i$, increases, that is, $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) \rightarrow \bar{x}_i$ as $i \rightarrow n$. In other words, as the number of observations increases, the relative importance of any specific observation value diminishes, but the overall importance of the average of all observations increases.

The conditional variance reduces to

$$V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) = \sigma^2 + \sigma^2\tau^2(\sigma^2 + i\tau^2)^{-1}. \tag{14}$$

Note that $V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ reduces monotonically towards σ^2 as the number of observations increase. In this application the posterior conditional variance decreases toward the variance of the ITF as i increases.

2.4 Case 1: $X \mid \mu, \sigma^2 \sim N_{dig}(\mu, 1)$ and $\mu \mid \sigma^2 \sim N_{dig}(0, 1)$

In the first example we shall consider the case where we have a one-stage prior mass function. The use of a one-stage prior mass function means that we feel our uncertain opinion about the conditional distribution of X can be fully specified by the moments that are described by one distribution, in this case the Normal. We choose our parameters to be $\mu_0 = 0$, $\sigma^2 = 1$ and $\tau^2 = 1$, so that

$$\begin{aligned} \mu \mid \sigma^2 &\sim N_{dig}(0, 1) && (15) \\ \text{and} \qquad X \mid \mu, \sigma^2 &\sim N_{dig}(\mu, 1). && (16) \end{aligned}$$

We identify realms $\mathbf{R}(X)$ and $\mathbf{R}(\mu)$ as having an interval width of 0.08. $\mathbf{R}(X)$ must cover every possible value that could realistically be generated as a measurement for X. We shall characterise the standard measurements of X as fully between -4 and 4. We shall also include a few extreme values ($\pm 5, \pm 6, \pm 7, \pm 8$) to represent possible extreme measurements. Thus the size of $\mathbf{R}(X)$ is 109.

Extreme measurements are often measured more crudely, either because the measurement device is not calibrated to record extreme observations with the same degree of fineness as it is for commonly observed values, or because the researchers may not regard measurement precision to be as important for less common observation values. An example of this situation is described in Ware and Lad (2003), where it is not possible to measure extreme observations to the normal level of fineness.

The elements of $\mathbf{R}(\mu)$ should cover every value of the mixing location parameter μ. In this case the realm will have a minimum value of -2 and a maximum value of 2, with the same spacing in the grid of possibilities as for X, 0.08. Thus the size of $\mathbf{R}(\mu)$ is 51.

We shall study the closeness of the conjugate Normal approximation to the precise computation based on the digitised Normal distribution under three different observational scenarios. The 'observed' X values shall be sampled from mass functions corresponding to the digitised $N(0, 1)$, $N(2, 1)$ and $N(4, 1)$ densities. When we sample from mass function $N_{dig}(0, 1)$ we expect the mode of our sampled values to be equal to the mode of the prior location mixing mass function, $\mu = 0$. When we sample from mass function $N_{dig}(2, 1)$, we obtain a sequence of observations that our prior mixing mass function suggests is unlikely. When we take a sample from mass function $N_{dig}(4, 1)$, we obtain a sequence of observations that our prior mixing mass function suggests is surprising.

2.4.1 Computation of $f(x \mid \mu, \sigma^2)$ and $f(\mu \mid \sigma^2)$

Our immediate aim is to form matrices $f(x \mid \mu, \sigma^2)$ and $f(\mu \mid \sigma^2)$. The number of rows in each matrix is equal to the size of $\mathbf{R}(X)$, and the number of columns is equal to the size of $\mathbf{R}(\mu)$. In this case the size of matrices $f(x \mid \mu, \sigma^2)$ and $f(\mu \mid \sigma^2)$ is 109×51. Matrix $f(x \mid \mu, \sigma^2)$ has rows corresponding to the elements of $\mathbf{R}(X)$ (i.e. from -8 to 8) and columns corresponding to elements of $\mathbf{R}(\mu)$ (i.e. from -2 to 2). We evaluate density $f(x \mid \mu, \sigma^2)$ at each grid-point, and then normalise the columns. Each column of $f(x \mid \mu, \sigma^2)$ is a mass function characterised as a digitised Normal density conditioned on the corresponding value of μ. For example the first column is $N_{dig}(-2, 1)$, the second column is $N_{dig}(-1.92, 1)$, and so on. A bar graph of each column shows that they look as if they are truncated digital Normal mass functions. Figure 1 shows two of these columns, corresponding to $N_{dig}(0, 1)$ and $N_{dig}(2, 1)$, as well as the bar graph of the mass function corresponding to a digitised $N(4, 1)$ density.

The bar graphs of the mass functions appear truncated because the range of $\mathbf{R}$(X) is less than the range of the non-zero values of the Normal densities that the mass functions are based on. A bar graph of the 26th column, $N_{dig}(0, 1)$, does look

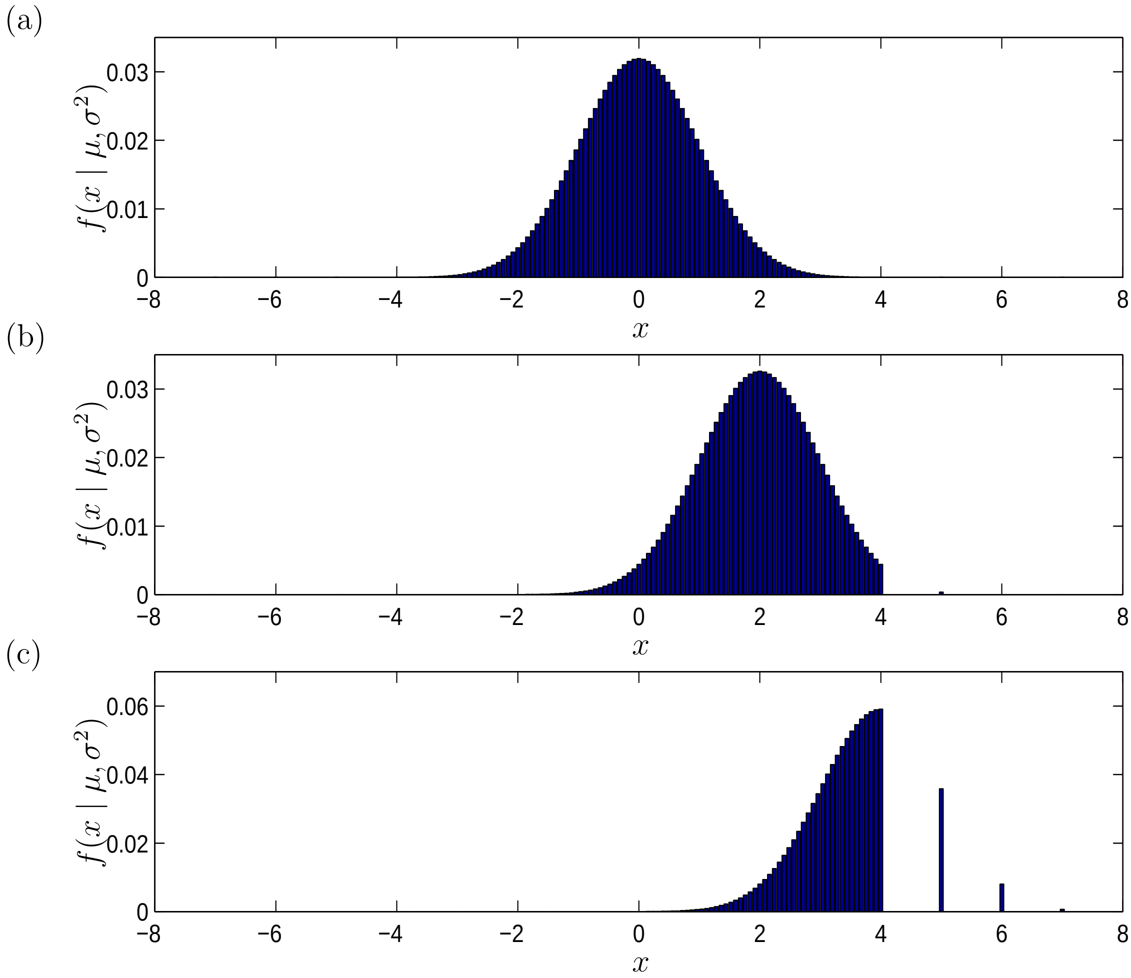

Figure 1: Bar graph of the digitised Normal mass functions evaluated at (a) $\mu = 0$, (b) $\mu = 2$ and (c) $\mu = 4$. Positive mass values are recorded for elements of $\mathbf{R}(X)$ only. Note that the scale along the y-axis in (c) is double that of (a) or (b).

as if it has a Normal shape because, although there are no points less than -4 or greater than 4 included in the computation of the mass function, these values have negligible mass when $\mu = 0$ and $\sigma^2 = 1$. A bar graph of the 51st column looks obviously truncated because $\mathbf{R}$(X) only has elements fully recorded on the interval from -4 to 4. Observe there is positive mass placed on $X = 5$ (see Figure 1(b)). This effect is demonstrated even more clearly by Figure 1(c), a bar graph of $N_{dig}(4, 1)$, where it appears the mass function has only been fully recorded on half its range. There are extreme elements with noticeable positive mass recorded at $X = 5, 6, 7$.

The prior mixing mass function is characterised as $N_{dig}(0, 1)$. To form matrix

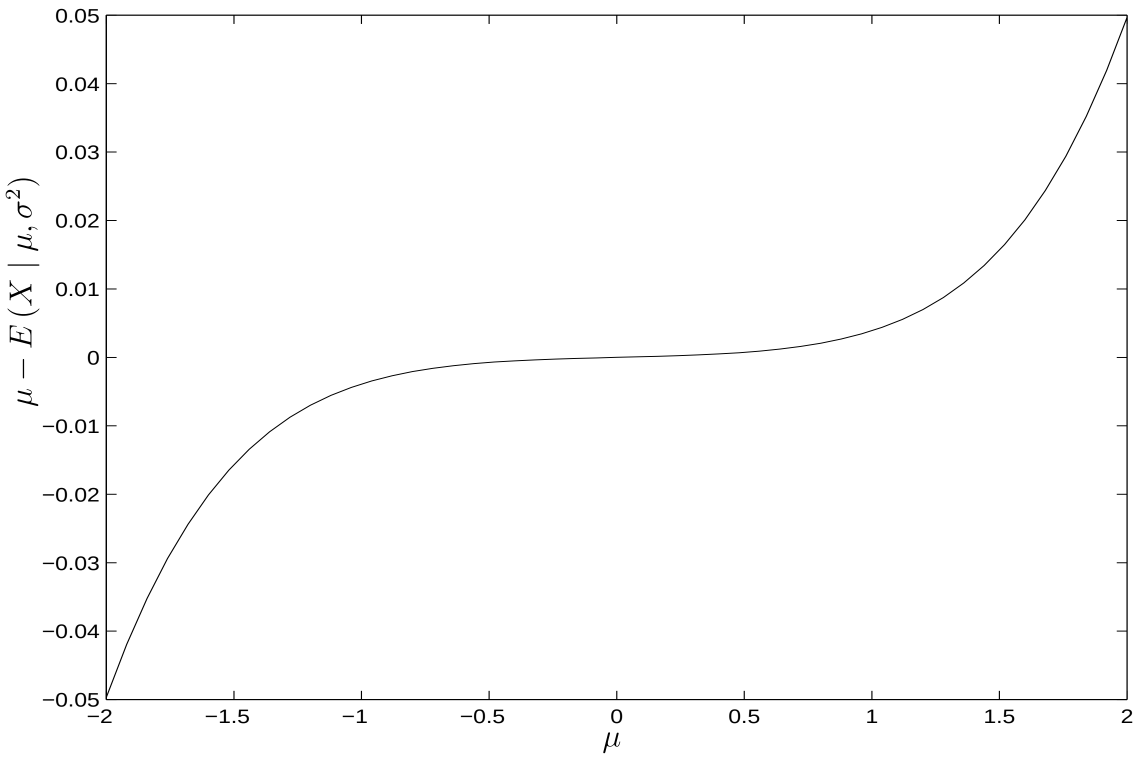

Figure 2: Magnitude of error in $\mathbf{E}(X \mid \mu, \sigma^2)$ using continuous conjugate methods rather than actual digital computations. Note that these errors are calculated directly from $f(x \mid \mu, \sigma^2)$.

$f(\mu \mid \sigma^2)$ evaluate the standard Normal density for each constituent element of $\mathbf{R}(\mu)$, and normalise. We now have $f(\mu \mid \sigma^2)$, a 1×51 vector. Replicate and tile this vector so that it becomes a matrix with 109 rows.

2.4.2 Computation of Conditional Moments for $f(x \mid \mu, \sigma^2)$

To compute conditional moments for the columns of $f(x \mid \mu, \sigma^2)$ we use the formulae,

$$E(X \mid \mu, \sigma^2) = \sum_x x f\left(x \mid \mu, \sigma^2\right) \tag{17}$$

$$\text{and} \qquad V\left(X \mid \mu, \sigma^2\right) = E(X^2 \mid \mu, \sigma^2) - \left[E\left(X \mid \mu, \sigma^2\right)\right]^2. \tag{18}$$

The expected value for each column of matrix $f(x \mid \mu, \sigma^2)$ can be computed using Equation 17. The expected values are symmetric about 0 and range from near -2, for $E(X \mid \mu = -2)$, to near 2, for $E(X \mid \mu = 2)$. Figure 2 shows the magnitude of error in $E(X \mid \mu, \sigma^2)$ when using continuous conjugate approximations rather than the actual computed values obtained using digital mass functions.

The calculation of $E(X \mid \mu = 2)$ produces a value of 1.95 (2dp) compared to the continuous conjugate approximation of $E(X \mid \mu = 2) = 2$. This is because, although $\mathbf{R}(X)$ is symmetric, the 51st column of the $f(x \mid \mu, \sigma^2)$ matrix, $f(x \mid \mu = 2, \sigma^2 = 1)$ is not. The 51st column of $f(x \mid \mu, \sigma^2)$ is symmetric about μ within the range $[0, 4]$. Beyond this range there are more components of $\mathbf{R}(X)$ below 0 than above 4. Consequently, the positive mass on $f(X < 0 \mid \mu = 2)$ is greater than the mass on $f(X > 4 \mid \mu = 2)$, and so $E(X \mid \mu = 2) < 2$. In contrast the conjugate approximation specifies complete symmetry about $\mu = 2$. Thus when Equation 17 is computed the actual digital calculation is smaller than the conjugate approximation. In fact the conjugate approximation will always be larger absolutely than the digital computation, except for $E(X \mid \mu = 0)$, when the conjugate approximation and digital computation are equal. Figure 2 illustrates that as $|\mu|$ decreases towards 0, $|E(X \mid \mu, \sigma^2) - \mu|$ decreases. This is because as μ decreases the range of values in $\mathbf{R}(X)$ that are symmetric about μ increase.

Values of $V(X \mid \mu, \sigma^2)$ range even more widely over the possible values of μ, as is shown in Figure 3. Whereas the conjugate continuous specification is of a variance constant over the different values of μ, the digitised conditional moments have smaller variances when located about μ values away from 0. As with the conditional expectation, this is because, as we move away from 0, the number of X values: (a) for which the mass of $N(\mu, 1)$ is non-negligible, and (b) are not included in $\mathbf{R}(X)$, increases.

2.4.3 Case 1A: Observations Generated from $\mathbf{X}_N \sim N_{dig}(0, 1)$

Case 1A: Digital Mass Functions. Now that matrices $f(x \mid \mu, \sigma^2)$ and $f(\mu \mid \sigma^2)$ have been computed we are in position to consider the three examples. We begin with observations generated from $N_{dig}(0, 1)$. We generated 25 observations using the procedure described in Section 2.1.3. In this case the mean of the prior mixing function is equal to the mean of the mass function we used to generate our observations.

We want to find the series of items $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$. Section 2.2 has detailed how this process can be undertaken: use Bayes' Theorem to update the prior mixing function, compute the updated predictive mass function and

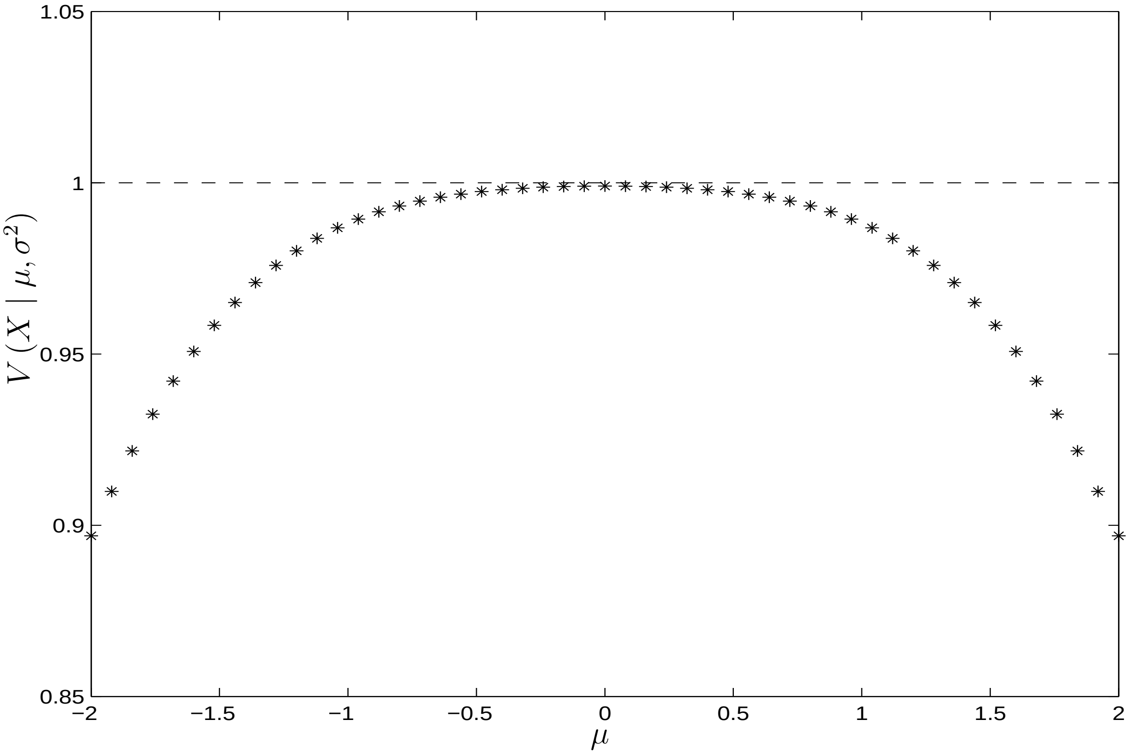

Figure 3: Values of $\mathbf{V}(X \mid \mu, \sigma^2)$ obtained using digital computation (marked by "$*$") and conjugate approximation (marked by "$--$"). Note that continuous conjugate theory specifies constant variance over all values of μ, while digital computations show the variance is dependent on μ.

then calculate the conditional moments. This process can be repeated as many times as required, which in this case is until we have found values for $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ for $i = 0, \ldots, 24$.

Case 1A: Conjugate Theory. A study of Equation 10 shows that the distribution of $\mathbf{X}_N$ can be written in the form

$$\mathbf{X}_N \sim N\left(\begin{array}{c} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{array}, \begin{bmatrix} 2 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 2 & & 1 \\ \vdots & \vdots & & \ddots & \\ 1 & 1 & 1 & & 2 \end{bmatrix}\right). \qquad (19)$$

It follows from Equations 13 and 14 that the posterior mean of the $(i+1)th$ observation, conditional on observing the first i observations is

$$E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) = \frac{\sum\limits_i x_i}{i+1}. \tag{20}$$

The posterior variance of the $(i+1)th$ observation, conditional on observing the first i observations is

$$V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) = \frac{i+2}{i+1}. \tag{21}$$

Case 1A: Results. Figure 4 shows the result of a typical example for the case where 25 observations are generated via a digitised $N(0,1)$ density. In this case the prior mixing function is the same as the conditional distribution of observations, $\mathbf{X}_{25}$, so it is not surprising that the conditional expectation of the digital computations and the conjugate approximations are very similar. The only slight disparity occurs for the first few observations. The conditional expectations are plotted in the top panel of Figure 4.

The two different conditional variances we calculate are very similar for $i \geq 5$. The first few conditional variances have larger values for the conjugate approximation than they do for the digital computation. The value of $V(X_1)$ is 1.760, but conjugate theory approximates it as 2. If $\mathbf{R}(\mu)$ had have been wider, that is, contained more extreme elements, then we could reasonably expect the digital mass value and conjugate approximation to be closer. As i increases the two types of conditional variance modelled become more similar.

We have just investigated the case where the conditional distribution of observations is the same as the mode of the prior mixing mass function. We have seen that the actual digital mass values of conditional moments, and the estimates obtained through conventional conjugate theory for the same conditional moments, are very similar.

2.4.4 Case 1B: Observations Generated from $\mathbf{X}_N \sim N_{dig}(2,1)$.

Case 1B: Digital Mass Functions. Now we shall consider a case where observations are generated from $\mathbf{X}_N \sim N_{dig}(2,1)$, where 2 is one of the endpoints of $\mathbf{R}(\mu)$. We construct simulated observation vector $\mathbf{X}_N$ in a similar way to Case 1A,

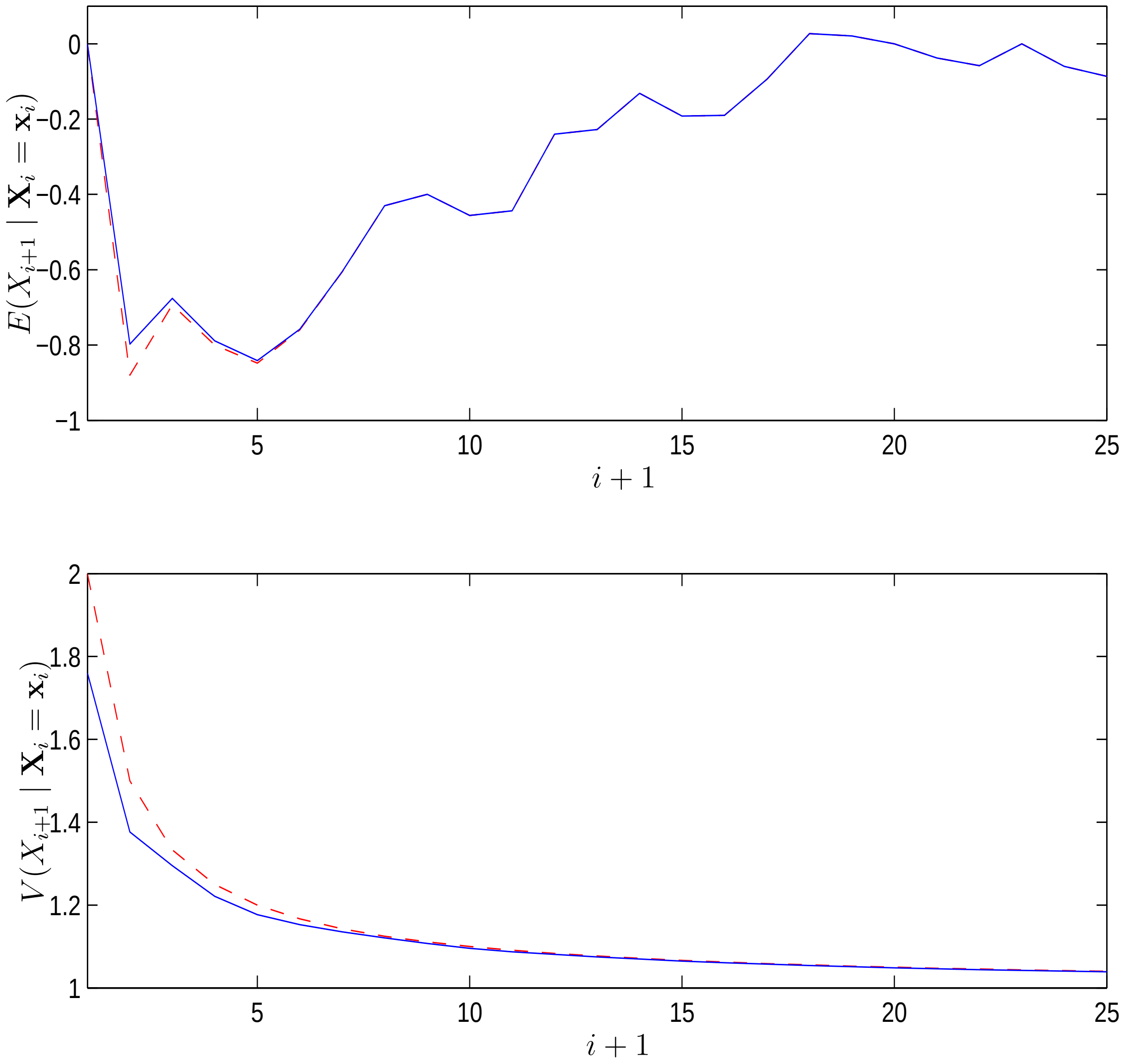

Figure 4: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{25} \sim N_{dig}(0, 1)$.

the difference being that the conditional distribution of observations is now characterised as a digital Normal density with parameters $\mu = 2$ and $\sigma^2 = 1$. In this case we observe 100 values of X, rather than 25, because both $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ take longer to stabilise when $\mu = 2$. We find a series of posterior mass functions, predictive mass functions and items $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ as described in Case 1A.

Case 1B: Conjugate Theory. A study of Equation 10 shows that the distribution of $\mathbf{X}_{100}$ can be written in the form $\mathbf{X}_{100} \sim N_{100}\left(\mathbf{2}, \mathbf{I}_{100} + \mathbf{1}_{100,100}\right)$. It follows

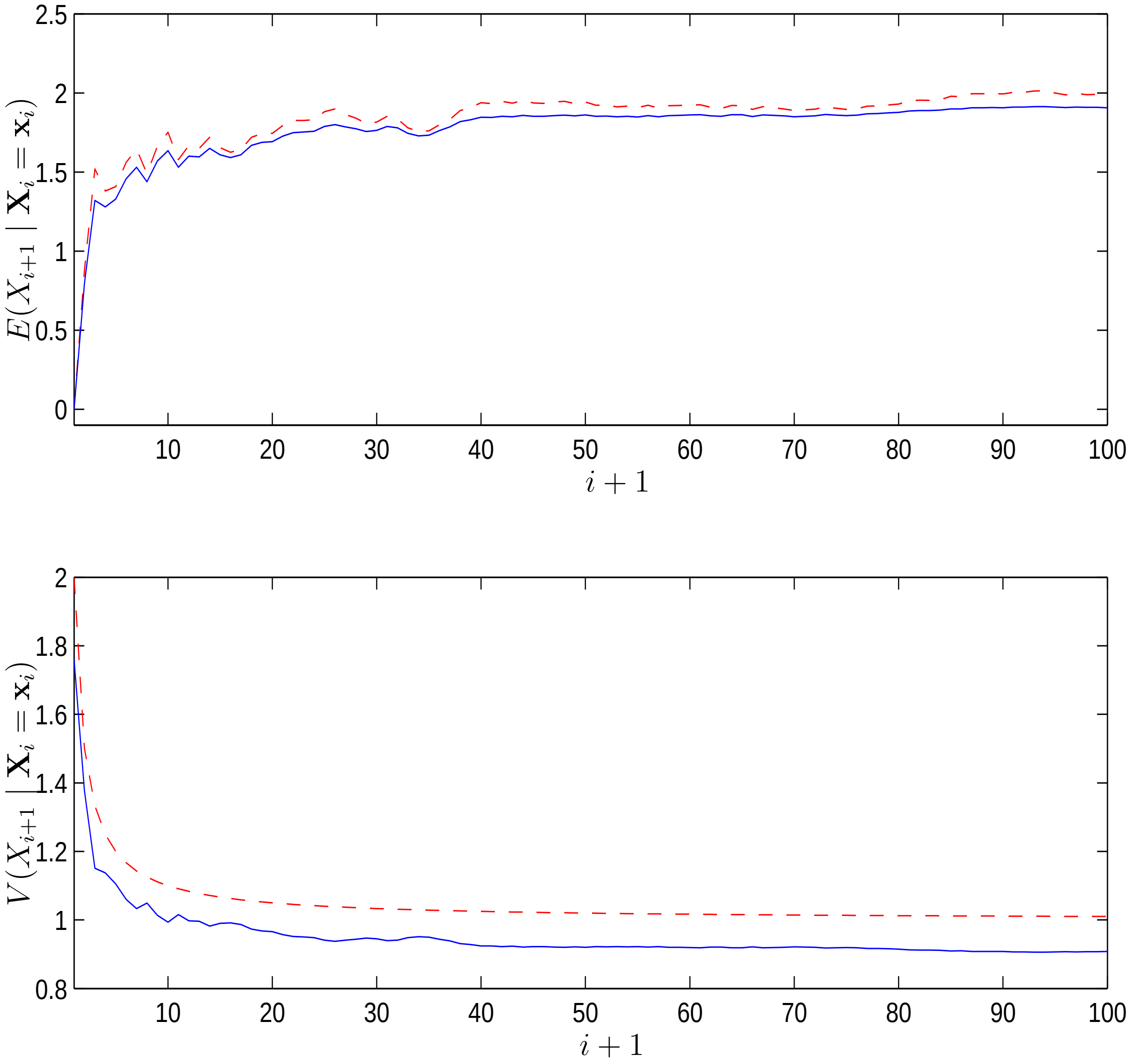

Figure 5: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{100} \sim N_{dig}(2,1)$.

from Equations 13 and 14 that both the posterior mean and variance of the $(i+1)th$ observation, conditional on observing the first i observations, are the same as in Case 1A, viz:

$$E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) = \frac{\sum_i x_i}{i+1} \tag{22}$$

and

$$V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) = \frac{i+2}{i+1}. \tag{23}$$

Case 1B: Results. The conditional expectation of $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ is plotted in the upper panel of Figure 5 for a typical example when 100 observations are generated from a digitised $N(2,1)$ distribution. The equivalent conditional variance is shown in the lower panel of Figure 5. The generating value of μ has been selected as one of the most extreme possibilities of $\mathbf{R}(\mu)$. There is a larger difference between the conditional expectations calculated using digital mass functions and conventional conjugate theory than in Case 1A. It is of the order of 0.1 relative to the actual expectation of approximately 1.9. The approximate values of $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ obtained by conjugate theory are always higher than the equivalent computed digital values. A corresponding observation can be made for the values of $V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$. The actual conditional variance is approximately 0.9 and the relative difference is about 0.1.

2.4.5 Case 1C: Observations Generated from $\mathbf{X}_N \sim N_{dig}(4,1)$.

Case 1C: Digital Mass Functions. Our final example in this Section considers the case where observations are generated from $N_{dig}(4,1)$. Notice that 4, the generating value of μ, is outside the range specified for $\mathbf{R}(\mu)$. The purpose of this example is to see how closely the continuous approximation and digital computation cohere when the generating value of μ has been selected to be far from our prior specification of μ. The observation vector is generated, and a series of updated mixing functions, updated predictive mass functions and items $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ are found, as in the two previous examples. In this case 250 observations were generated.

Case 1C: Conjugate Theory. A study of Equation 12 shows that the formulae for $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$, and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ are the same as in the previous two examples.

Case 1C: Results. A typical sequence of observations generated from a digitised $N(4,1)$ distribution are shown in Figure 6. As we expect, the most commonly selected elements of $\mathbf{R}(X)$ are those close to 4. In the sequence of 250 observations we only observe $(X_i > 4)$ on 9 occasions. Because the (non-digitised) $N(4,1)$ density is symmetric about 4 it is reasonable to expect that the number of times we observe $X_i = 3$ and $X_i = 5$, and the number of times we observe $X_i = 2$ and $X_i = 6$ will be

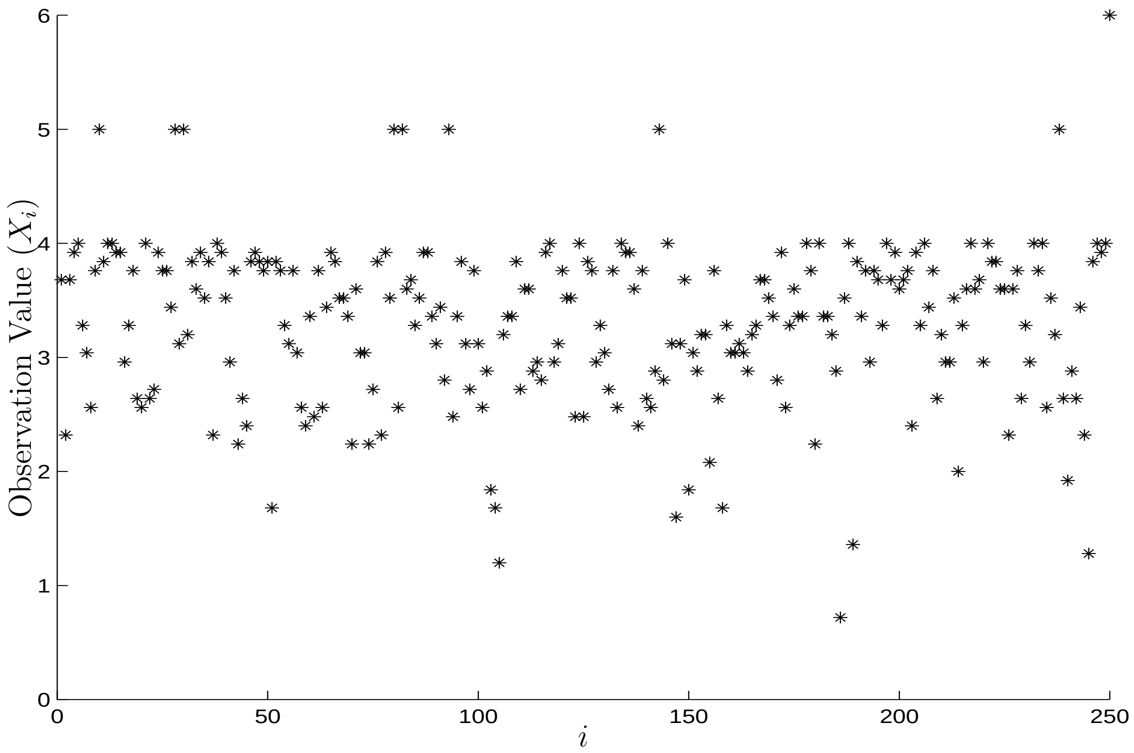

Figure 6: Timeplot demonstrating the sequence of observations $\mathbf{X}_{250} \sim N_{dig}(4, 1)$.

approximately the same. In this case we observe $X_i = 3$ on 9 occasions compared to $X_i = 5$ on 8 occasions.We observe $X_i = 2$ and $X_i = 6$ once each.

Figure 7 shows the conditional expectation and variance for a typical example when we have 250 observations. $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ is plotted in the upper panel of Figure 7. It shows a large difference between the conjugate approximation (which is approximately 3.3) and the actual digital computation ($\approx$ 1.95). The limit of the conjugate approximation approaches the arithmetic mean of the i observations as $i \to \infty$, so a value of $\approx$ 3.3 is not surprising considering the observed $\mathbf{X}_i$ (see Figure 6). The actual computed value of $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ will never be larger than 2, because the updated mixing function has positive mass only for elements of $\mathbf{R}(\mu)$. Since vector $f\left(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i, \sigma^2\right)$ has positive mass only from $[-2, 2]$, when $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ is computed the maximum value it can attain is 2.

The sequence of items $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ obtained when $\mathbf{X}_N \sim N_{dig}(4, 1)$ are similar to those obtained when $\mathbf{X}_N \sim N_{dig}(2, 1)$. As in Case 1B the conjugate approximation of $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ is larger that the digital computation by approxi-

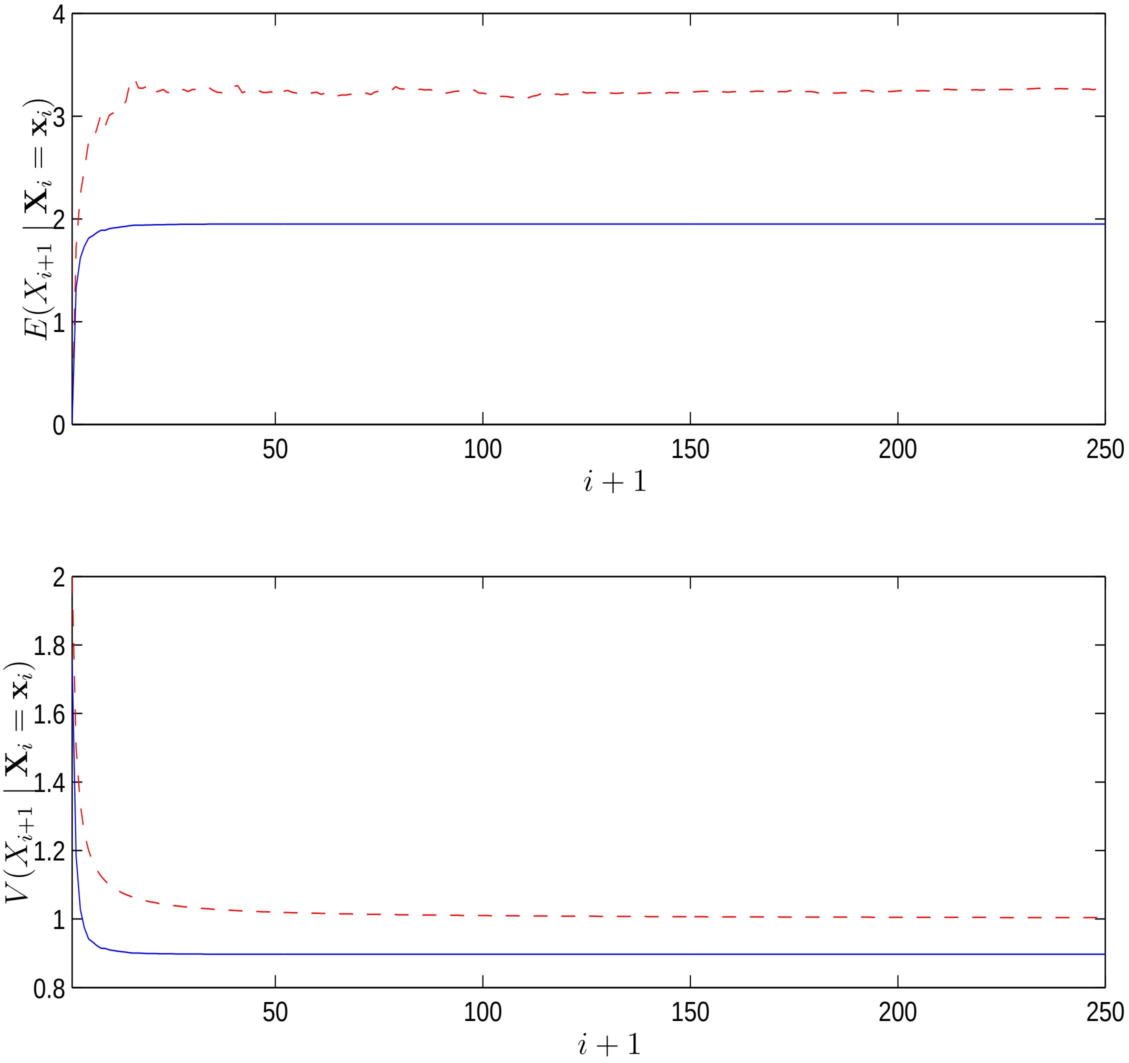

Figure 7: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{250} \sim N_{dig}(4, 1)$.

mately 0.1 relative to the real variance of 0.9. The similarity between the conditional variance for Case 1B and Case 1C is reflected in the similarity between the lower panels of Figure 5 and Figure 7.

We have seen that when recorded observations are similar to what the researcher expected they would be, the actual digital values of the conditional moments are similar to the conjugate approximation. When a researcher observes values they find surprising, the continuous approximations are some distance from the actual conditional moments. Next we shall investigate the similarity of posterior means and variances when a two-stage prior distribution is specified.

3 Approximation of Posterior Means and Variances when a Two-Stage Prior Distribution is Specified

So far in this Report we have considered how well conventional conjugate mixture-Normal theory can approximate real discrete mass functions, represented by digital Normal mass functions, when the prior distribution is fully specified. In this Section we shall extend this approach and investigate how well Normal-Gamma mixture-Normal conjugate theory approximates measurements obtained through the use of real discrete mass functions when they are representable by digital Normal mass functions and digital Gamma mass functions.

If we assert that the precision of the X observations, π, has a Gamma distribution with parameters α and β, the conditional distribution of μ, given π, is a Normal distribution with mean μ_0 and precision $\tau\pi$, and that X, conditioned on μ and π, is distributed Normally with mean μ and precision π. Then using an extension of Equation 1, the density for X can be derived by

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x \mid \mu, \pi) f(\mu \mid \pi) f(\pi) d\mu d\pi & (24) \\ &= \int_{-\infty}^{\infty} f(x \mid \mu) f(\mu) d\mu. & (25) \end{aligned}$$

Conventional conjugate theory tells us that, for a single case, X has a general t-distribution with location parameter μ_0, scale parameter $\tau\alpha/(1+\tau)\beta$ and shape parameter 2α. To summarise this situation symbolically,

$$\begin{aligned} \pi &\sim \Gamma(\alpha, \beta) & (26) \\ \mu \mid \pi &\sim N(\mu_0, \tau\pi) & (27) \\ X \mid \mu, \pi &\sim N(\mu, \pi) & (28) \\ \text{and} \qquad X &\sim t(2\alpha, \mu_0, \tau\alpha/(1+\tau)\beta). & (29) \end{aligned}$$

Under a Bayesian framework we say $f(\pi)$ is the prior density for π, $f(\mu \mid \pi)$ is the prior density for μ, $f(x \mid \mu, \pi)$ is the ITF and $f(x)$ is the predictive density for X. $f(\pi)$ and π are often called the hyperprior and hyperparameter respectively. This form of distribution is known as an hierarchical Bayes model.

In the previous section the one-stage prior, $f(\mu \mid \sigma^2)$, was used. That is, we assumed the spread was known, and had the value σ^2, but that the location of the prior was unknown, and was distributed Normally with parameters μ_0 and τ^2. Now we are including an additional level of prior modeling by assuming that both the location, conditioned on the spread and represented by μ, and the spread, represented by π, of the observations are unknown, and placing a prior distribution on each of them. We assess the distribution of $\mu \mid \pi$ as in Case 1, but now we also need to assess the distribution of π. This is an example of a two-stage prior distribution.

Notice that a major difference in this Section is that we now parameterise the Normal distribution by its precision, π, rather than its variance, $\sigma^2 \equiv \pi^{-1}$. The only time in the remainder of Section 3 that we shall refer to variance is in the calculation of item $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$.

3.1 Discrete Mass Functions Characterised as Digitised Normal and Digitised Gamma Densities

This Section will involve working with probability mass functions $f(x \mid \mu, \pi)$ and $f(\mu \mid \pi)$, which are characterised as digitised Normal densities, and $f(\pi)$, which is characterised as a digitised Gamma density. To compute the predictive mass function of X we follow a procedure similar to that described in Section 2. In that case we worked with the mass functions $f(x \mid \mu, \sigma^2)$ and $f(\mu \mid \sigma^2)$, and created two matrices of equal dimensions. Now we need to consider three mass functions. To account for the extra mass function we shall represent $f(x \mid \mu, \pi)$, $f(\mu \mid \pi)$ and $f(\pi)$ in array form. The three arrays constructed will be of equal size

Before we can form the three arrays needed to undertake the required computations we must define realms for X, μ and π.

3.1.1 Forming Arrays of Mass Values

To generate the digitised space we are interested in, we first identify realms $\mathbf{R}(X)$, $\mathbf{R}(\mu)$ and $\mathbf{R}(\pi)$, which represent the possible measurement values of X, and mixing possibilities for μ and π respectively. The specification of $\mathbf{R}(X)$ and $\mathbf{R}(\mu)$ has been discussed previously. $\mathbf{R}(\pi)$ will represent the spread of the digitised mixture function with parameter π, and should include every value of π that could be relevant to its subject.

After identifying the three realms of interest, we form an array of mass values for $f(x \mid \mu, \pi)$. The array will have height corresponding to the candidate X values, width corresponding to the candidate μ values and depth corresponding to the possible values for π. Evaluate density $f(x \mid \mu, \pi)$ at each grid-point. Normalise the columns for later use, when they will aid in the computation of the posterior mass function $f(\mu \mid X = x, \pi)$.

Arrays $f(\mu \mid \pi)$ and $f(\pi)$ also have height corresponding to the size of $\mathbf{R}(X)$, width corresponding to the size of $\mathbf{R}(\mu)$ and depth corresponding to the size of $\mathbf{R}(\pi)$. To construct $f(\mu \mid \pi)$, form a matrix whose dimensions correspond to candidate μ and π values. Evaluate density $f(\mu \mid \pi)$ at each grid-point. Replicate and tile the matrix until it is the same size as $f(x \mid \mu, \pi)$. Note that every plane corresponding to a member of $\mathbf{R}(X)$ will be identical.

To construct array $f(\mu)$ a similar method is used. First, evaluate $f(\pi)$ for each element of $\mathbf{R}(\pi)$. The resulting vector represents a mass function characterised as a digitised Gamma density. Replicate and tile the vector along both the X and μ dimensions. Values in this array will vary in the π dimension only. All three arrays will now be the same size.

3.1.2 Generating Sequences of Observation Values

A vector of length N containing observation values is generated by following the sequence of steps outlined in Section 2.1.3. The only difference is that the mass function the X's are drawn from is parameterised by μ and π, rather than by μ and σ^2. To this point we have constructed realms $\mathbf{R}(X)$, $\mathbf{R}(\mu)$ and $\mathbf{R}(\pi)$, arrays $f(x \mid \mu, \pi)$, $f(\mu \mid \pi)$ and $f(\pi)$ and $\mathbf{X}_N$, a vector containing N observation values.

3.2 Computing Posterior Means and Variances of Posterior Discrete Densities

To find exact values for $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ for the discrete mass functions characterised as digitised Normal and digitised Gamma densities, we compute

$$f\left(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) = \sum_{\mu}\sum_{\pi} f\left(x_{i+1} \mid \mu, \mathbf{X}_i = \mathbf{x}_i, \pi\right) f\left(\mu, \pi \mid \mathbf{X}_i = \mathbf{x}_i\right), \tag{30}$$

which is an extension of Equation 6. The first function in this sum of products does not change. It is always represented by array $f(x \mid \mu, \pi)$. The second function is a combination of Equations 26, 27 and 28. After each observed value of $(\mathbf{X}_i = \mathbf{x}_i)$, it is updated via Bayes' Theorem, a consequence of which is

$$f(\mu, \pi \mid \mathbf{X}_i = \mathbf{x}_i) = \frac{f(\mathbf{X}_i = \mathbf{x}_i \mid \mu, \pi) f(\mu, \pi)}{\sum_{\mu}\sum_{\pi} f(\mathbf{X}_i = \mathbf{x}_i \mid \mu, \pi) f(\mu, \pi)} \qquad \text{for} \quad i = 1, 2, \ldots, N-1. \tag{31}$$

Before we can implement Bayes' Theorem we must calculate the joint mass function for (μ, π). It is widely known that

$$f(\mu, \pi) = f(\mu \mid \pi) f(\pi). \tag{32}$$

Thus to obtain $f(\mu, \pi)$ we merely multiply array $f(\mu \mid \pi)$ element-wise with array $f(\pi)$. Once $f(\mu, \pi)$ is found we follow similar steps to those outlined in Section 2.2 on page 9:

1. Observe $X_i = x_i$.

2. Extract the matrix corresponding to X_i from array $f(x \mid \mu, \pi)$. This is the ITF through μ and π from $X_i = x_i$ to X_{i+1}.

3. Implement Bayes' Theorem and thus obtain the updated mixing function, matrix $f(\mu, \pi \mid \mathbf{X}_i = \mathbf{x}_i)$. Replicate and tile this matrix so it has the same size as array $f(x \mid \mu, \pi)$.

4. Calculate $f(x_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ by Equation 30.

5. Compute $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$.

Repeat steps 1–5 as many times as required.

3.3 Conventional Conjugate Normal-Gamma Mixture-Normal Theory

A multivariate Normal distribution that treats the components of $\mathbf{X}$ independently can be identified by parameters μ and π. The density function for a vector of quantities $\mathbf{X} \in \mathcal{R}^K$ is

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (2\Pi)^{-K/2} \, \pi^{K/2} \mathrm{exp} \left[-\pi \sum_{i=1}^{K} (x_i - \mu)^2 /2 \right] dM\,(\mu, \pi)\,, \tag{33}$$

where Π denotes the real number pi. We use the capital letter Π for the real number pi, rather than the usual lower case letter, in an attempt to avoid confusion with the mixing parameter $\pi = (\sigma^2)^{-1}$. We say $f(\mathbf{x})$ is a mixture of conditionally independent Normal densities with mixing parameters μ and π and mixing distribution function M, and it is denoted by $\mathbf{X} \sim M\text{-}N_K\,(\mu, \pi)$.

The product of a conditional density function and a marginal density is a joint density function. If the conditional density function, denoted $(\mu \mid \pi)$, is Normal with parameters μ_0 and $\tau\pi$, and if the marginal density, π, is Gamma with parameters α and β, then the joint density function for (μ, π) is

$$f(\mu, \pi) \propto (\tau\pi)^{1/2} \,\mathrm{exp} \left[-\tau\pi\,(\mu - \mu_0)^2 /2 \right] \pi^{\alpha-1} \mathrm{exp}\,(-\beta\pi) \qquad ((\mu, \pi) \in (\mathcal{R}, \mathcal{R}^+)), \tag{34}$$

and we say the joint density is a member of the Normal-Gamma family of distributions, denoted $(\mu, \pi) \sim N\Gamma\,(\mu_0, \tau, \alpha, \beta)$.

Suppose the components of $\mathbf{X}$ are regarded exchangeably and that $\mathbf{X} \sim M\text{-}N_K(\mu, \pi)$ with $M(\mu, \pi)$ specified as $N\Gamma\,(\mu_0, \tau, \alpha, \beta)$. If $\mathbf{X}$ is partitioned into $\mathbf{X}_1$ and $\mathbf{X}_2$, of sizes K_1 and K_2 respectively, then $\mathbf{X}_2 \mid (\mathbf{X}_1 = \mathbf{x}_1) \sim M\text{-}N_{K_2}\,(\mu, \pi)$, with the conditional mixing function $M(\mu, \pi \mid \mathbf{X}_1 = \mathbf{x}_1)$ in the Normal-Gamma form

$$N\Gamma \left[\frac{\tau\mu_0 + K_1 \bar{x}_{K_1}}{\tau + K_1}, \tau + K_1, \alpha + \frac{K_1}{2}, \beta + \left(\frac{K_1}{2}\right) s_{K_1}^2 + \frac{\tau K_1 \left(\bar{x}_{K_1} - \mu_0\right)^2}{2\,(K_1 + \tau)} \right], \tag{35}$$

where $\bar{x}_{K_1}$ is the arithmetic mean of $\mathbf{X}_1$ and $s_{K_1}^2$ is the average squared difference, $K_1^{-1} \sum_{i=1}^{K_1} (x_i - \bar{x})^2$. For details see the text of Lad (1996, pp. 395–397, 408–412).

In this application involving the forecast of X_{i+1} given $(\mathbf{X}_i = \mathbf{x}_i)$, we are interested in the case when $\mathbf{X}_1 = (\mathbf{X}_i = \mathbf{x}_i)$ and $\mathbf{X}_2 = X_{i+1}$. The conditional expectation reduces to

$$E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i) = \frac{\tau\mu_0 + i\bar{x}_i}{\tau + i}. \tag{36}$$

Initially the conditional expectation depends solely on μ_0, but as i increases the weighting given to μ_0 steadily decreases. $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ will be increasingly strongly influenced by the arithmetic mean of the observed quantities. As more observations are recorded each individual observation, and the prior mean, will be less influential, but the arithmetic mean of the observed quantities will be more influential.

The conditional variance is

$$V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i) = \frac{(\tau + i + 1)\left[2\beta + is_i^2 + \tau i\left(\bar{x}_i - \mu_0\right)^2 / (i + \tau)\right]}{(\tau + i)\left[2\alpha + i - 2\right]}. \tag{37}$$

The conditional variance is initially equal to the prior variance, $\beta(1+\tau)/\tau(\alpha - 1)$. As $i \to \infty$, $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i) \to s_i^2$, where s_i^2 is the average squared difference of the conditioning observations from their mean. That is, as the number of observations increases, each individual observation becomes less important, but the overall importance of s_i^2 increases in determining the predictive variance for the next quantity X_{i+1}.

3.4 Case 2: $\pi \sim \Gamma_{dig}(2,2)$, $\mu \mid \pi \sim N_{dig}(0,\pi)$ and $X \mid \mu, \pi \sim N_{dig}(\mu, \pi)$

The examples in this Section consider the case where we have a digitised two-stage prior mass function. We shall study the closeness of the conjugate Normal-Gamma mixture-Normal approximation to the precise computation based on the digitised Gamma and digitised Normal distributions under different scenarios. Parameters are chosen to be $\alpha = 2$, $\beta = 2$, $\mu_0 = 0$ and $\tau = 1$, so that

$$\pi \sim \Gamma_{dig}(2,2) \tag{38}$$

$$\mu \mid \pi \sim N_{dig}(0,\pi) \tag{39}$$

$$\text{and} \qquad X \mid \mu, \pi \sim N_{dig}(\mu, \pi). \tag{40}$$

The 'observed' X's shall be sampled from mass functions corresponding to digitised Normal distributions with parameters $\mu = 0, 2, 4$ and $\pi = 0.25, 0.4, 1$. Note that the μ values we shall use are the same as in Case 1.

We identify realms $\mathbf{R}(X)$, $\mathbf{R}(\mu)$ and $\mathbf{R}(\pi)$ as having as interval width of 0.08. We characterise the standard measurements of X as fully between -6 and 6, with summary extreme values ($\pm 7, \pm 8, \pm 9, \pm 10$) to represent possible extreme measurements. The size of $\mathbf{R}(X) = 159$. The realm of μ has the same endpoints as in Case 1, -2 and 2, meaning the size of $\mathbf{R}(\mu)$ is 51. $\mathbf{R}(\pi)$ should cover every reasonable value of interest for the mixing spread parameter π. In this case the realm of π will have a minimum value of 0.4 and a maximum value of 2. Thus the size of $\mathbf{R}(\pi)$ is 21.

3.4.1 Computation of $f(x \mid \mu, \pi)$, $f(\mu \mid \pi)$ and $f(\pi)$

Before we can compute any of our desired items we must form arrays $f(x \mid \mu, \pi)$, $f(\mu \mid \pi)$ and $f(\pi)$. All three arrays will be size $159 \times 51 \times 21$. The only array whose entries are all distinct is $f(x \mid \mu, \pi)$. Entries along its height, which we call the X-axis, correspond to elements of $\mathbf{R}(X)$. Elements along its width, the μ-axis, correspond to the elements of $\mathbf{R}(\mu)$ and elements along its depth, the π-axis, correspond to the elements of $\mathbf{R}(\pi)$. A combination of any single element from $\mathbf{R}(\mu)$ and any single element from $\mathbf{R}(\pi)$ will have a related vector of mass values, corresponding to each element of $\mathbf{R}(X)$. Each of these mass functions is characterised as a digitised $N(\mu, \pi)$ density. For example, if the value of μ is chosen to be -2 and the value of π is chosen to be 0.4, then the corresponding mass function, denoted by the values along the X-axis, is characterised as a digitised Normal density with mean -2 and precision 0.4. If we move one place along the μ-axis, to $\mu = -1.92$, the corresponding mass function is characterised as a digitised Normal density with mean -1.92 and precision 0.4.

To construct array $f(\mu \mid \pi)$, first form a matrix of size 51×21. The values along the μ-axis correspond to elements of $\mathbf{R}(\mu)$. This has length 51. Values along the π-axis correspond to elements of $\mathbf{R}(\pi)$. This has length 21. At each gridpoint evaluate the mass of a $N(0, \pi)$ distribution at value μ. Normalise along the dimension corresponding to $\mathbf{R}(\mu)$. Replicate and tile the matrix along the X-axis until it has height 109.

Array $f(\mu)$ is constructed from a mass function which corresponds to a digitised Gamma$(2, 2)$ density. The mass function is evaluated for each member of $\mathbf{R}(\pi)$, and is then replicated and tiled along both the X-axis and μ-axis.

3.4.2 Case 2A: Observations Generated from $\mathbf{X}_N \sim N_{dig}(0,1)$

Case 2A: Digital Mass Functions Now that arrays $f(x \mid \mu, \pi)$, $f(\mu \mid \pi)$ and $f(\pi)$ have been computed we are in a position to consider our examples. We generate 500 observations from $N_{dig}(0,1)$ and find the sequence of items $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ by following the steps described in Section 3.2 for $i+1 = 1, \ldots, 500$.

Case 2A: Conjugate Theory It follows from Equations 36 that the posterior mean of the $(i+1)th$ observation, conditional on observing the first i observations, is

$$E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i) = \frac{i\bar{x}_i}{i+1}. \tag{41}$$

As i increases, the conditional expectation tends toward the arithmetic mean of $\mathbf{X}_i$.

Equation 37 implies that the posterior variance of the $(i+1)th$ observation, conditional on observing the first i observations, is

$$V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i) = \frac{4 + is_i^2 + i\bar{x}_i^2/\,(i+1)}{(i+1)}. \tag{42}$$

As the number of observations increase we expect $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ to tend to s_i^2.

Case 2A: Results. Figure 8 shows the result of a typical example when $\mathbf{X}_{500} \sim N_{dig}(0,1)$. The mode of the prior distribution is the same as the mode of the conditional distribution of observations $\mathbf{X}$. Consequently it is not surprising that the conditional expectations for the digital computations and conjugate approximation are almost identical for all values of i.

The two conditional variances are closely linked. Neither conditional variance decreases monotonically, unlike in Case 1. Notice how similar the fluctuations in $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ are. The difference between the digital computation and the conjugate approximation steadily decreases until they are almost identical whenever $i > 150$.

3.4.3 Case 2B: Observations Generated from $\mathbf{X}_N \sim N_{dig}(0, 0.4)$

Case 2B: Digital Mass Functions. The items required, $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ and $V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$, can be found in the manner outlined in Case 2A.

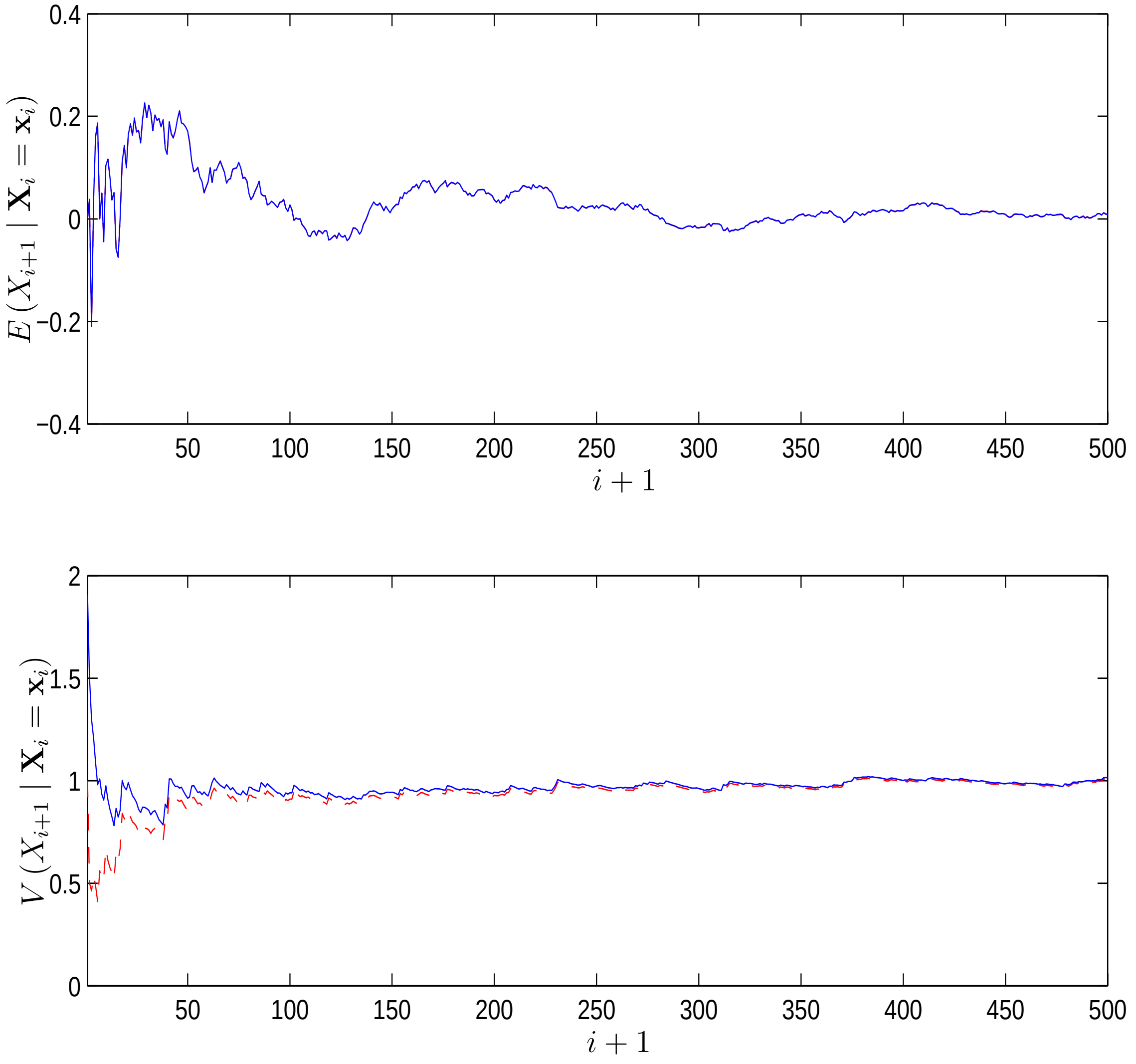

Figure 8: Conditional Expectations and Variances of the Digital Mass Function (blue line) and Conventional Conjugate Theory (red line) when $\mathbf{X}_{500} \sim N_{dig}(0,1)$. The conditional expectations are indistinguishable.

Case 2B: Conjugate Theory. A study of Equations 36 and 37 show the conditional posterior mean and variance remain the same as in Case 2A.

Case 2B: Results. Figure 9 shows the result of a typical example when $\mu = 0$, $\pi = 0.4$ (equivalent to $\sigma^2 = 2.5$) and $n = 5000$. As in Case 2A, the conditional expectations for the digital computations and conjugate approximation are almost identical for all values of i. Interestingly, the conjugate approximation of the conditional expectation appears to be slightly more unstable than the digital computation, most noticeably when $2000 < i$. Notice that the conjugate approximation fluctuates

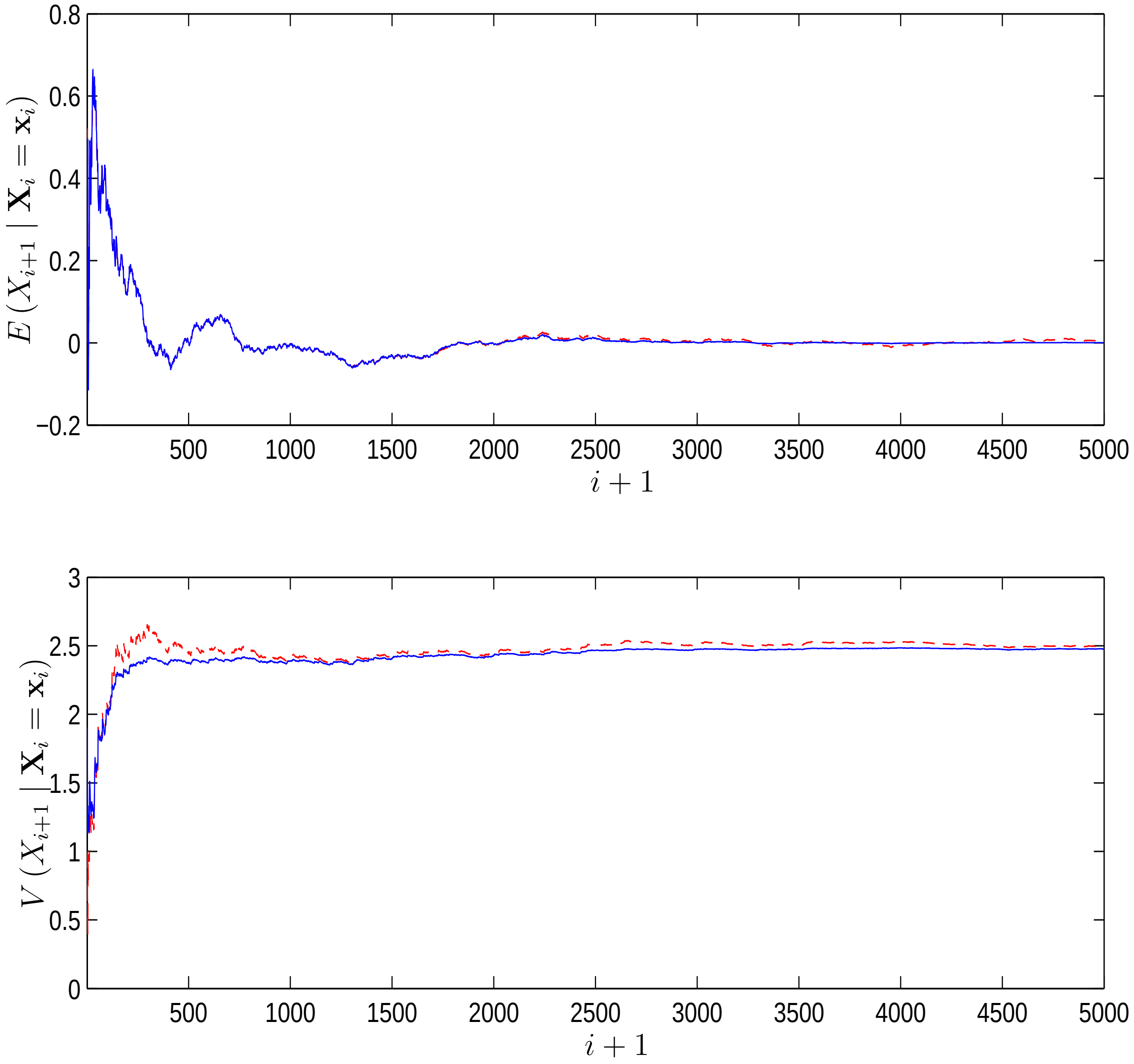

Figure 9: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{5000} \sim N_{dig}(0, 0.4)$.

slightly above and below the digital computation, which seems to be constant.

The two conditional variances are similar. Note that when $i < 1000$ the conjugate approximation fluctuates considerably more than the digital computation. The conditional variance of the digital computation appears to reach stability at $i \approx 3000$. The variance of the conjugate approximation is larger than the variance of the digital computation before stability is reached. The approximate variance decreases gradually until by $i \approx 4250$ the two conditional variances are very similar, nevertheless the conjugate approximation remains slightly larger.

3.4.4 Case 2C: Observations Generated from $\mathbf{X}_N \sim N_{dig}(0, 0.25)$

For Case 2C the observations are generated from a mass function characterised as a Normal density with mean equal to zero, which is also the value of the prior mean. The generating value of π is 0.25, which is outside the range of $\mathbf{R}(\pi)$. If a researcher, who had specified the same prior mixture mass function as we have, observed these values, they would notice that the observations were centred around 0, which is as they would expect, but would be surprised by how widely the data was spread.

The result of a typical example when $\mu = 0$, $\pi = 0.25$ and $n = 5000$ is shown in Figure 10. As in the previous two cases, the conditional expectation for the digital computation and conjugate approximation are almost identical for all values of i.

The two conditional variances are quite different. Due to the range of $\mathbf{R}(\pi)$, the conditional variance of the digital computation cannot take a value higher than 2.5. Remember that $\sigma^2 = 2.5$ is equivalent to $\pi = 0.4$. Although the observations are selected from a digitised Normal density with $\pi = 0.25$ ($\equiv \sigma^2 = 4$), the conditional variance of the conjugate approximation stays slightly below 4, $V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) \approx 3.8$. This is because $\mathbf{R}(X)$, the realm that contains all possible X values, is only fully specified on $[-6, 6]$. Thus the selection of $\mathbf{X}$ is likely to contain few values where $X < -6$ or $X > 6$. Consequently the 5000 observations sampled are likely to contain fewer members a long way away from the mean than they would if X was fully specified over a larger range. Thus the conjugate approximation is smaller than 4.

Suppose this example were to be repeated with observations drawn from a digitised Normal density with $\mu = 0$ and $\pi < 0.25$. By the above reasoning we expect that there will be an even larger difference between π, the approximated conditional variance, and the computed conditional variance.

3.4.5 Case 2D: Observations Generated from $\mathbf{X}_N \sim N_{dig}(2, 1)$

Figure 11 shows the result of a typical example when 500 observations are selected from a digitised Normal with $\mu = 2$ and $\pi = 1$. We observe results similar to those from Case 2A. After an initial period of instability the two conditional expectations are very similar. The conditional variances are also very similar, again after initial

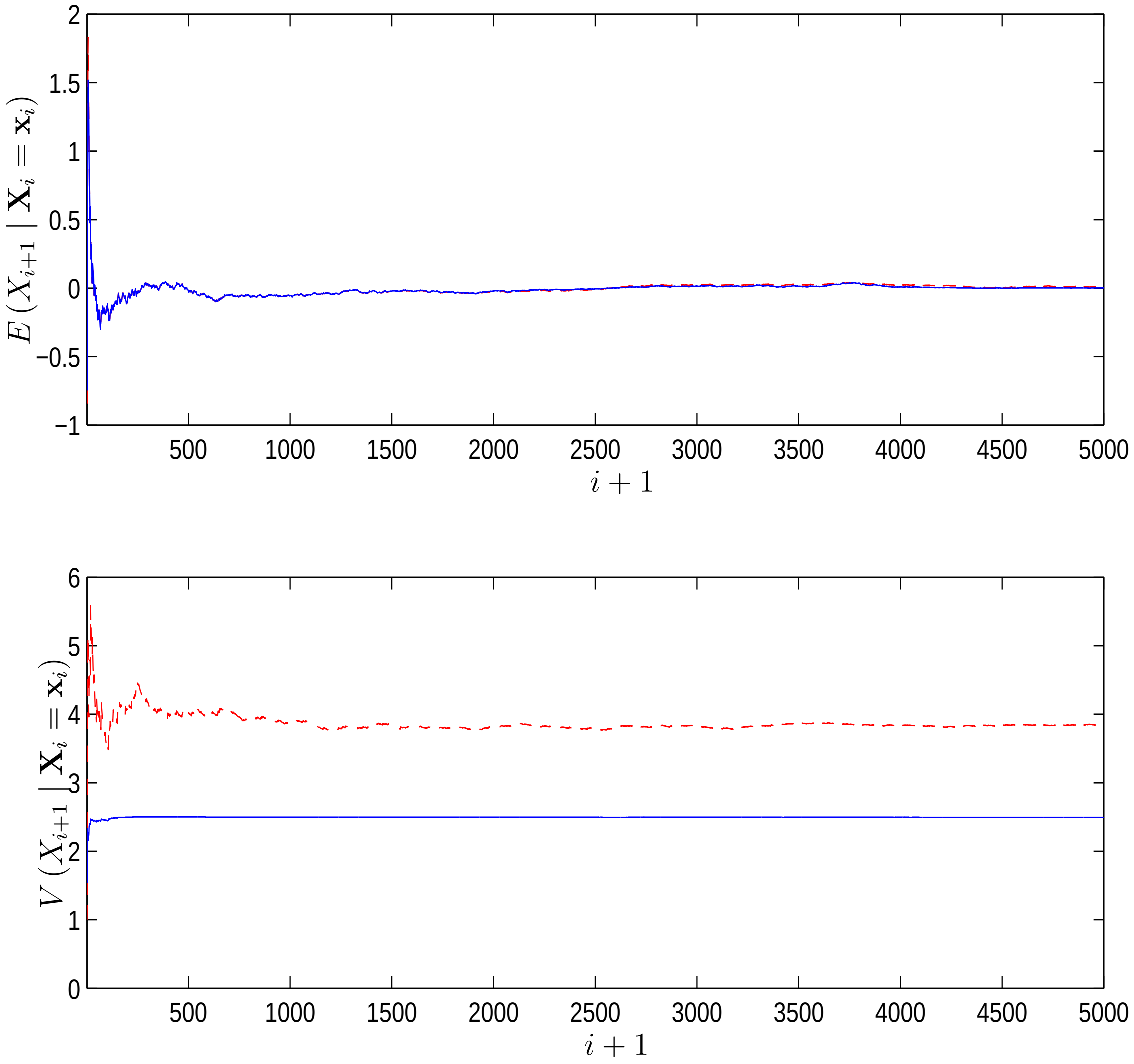

Figure 10: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{5000} \sim N_{dig}(0, 0.25)$.

instability.

3.4.6 Case 2E: Observations Generated from $\mathbf{X}_N \sim N_{dig}(2, 0.4)$

In Case 2E, 5000 observations are drawn from a mass function characterised as a digital $N(2, 0.4)$ density. Figure 12 shows the conditional expectations stabilise to the same value. Again the conjugate approximation is more unstable than the digital computation.

The conjugate approximation of the conditional variance is slightly larger than

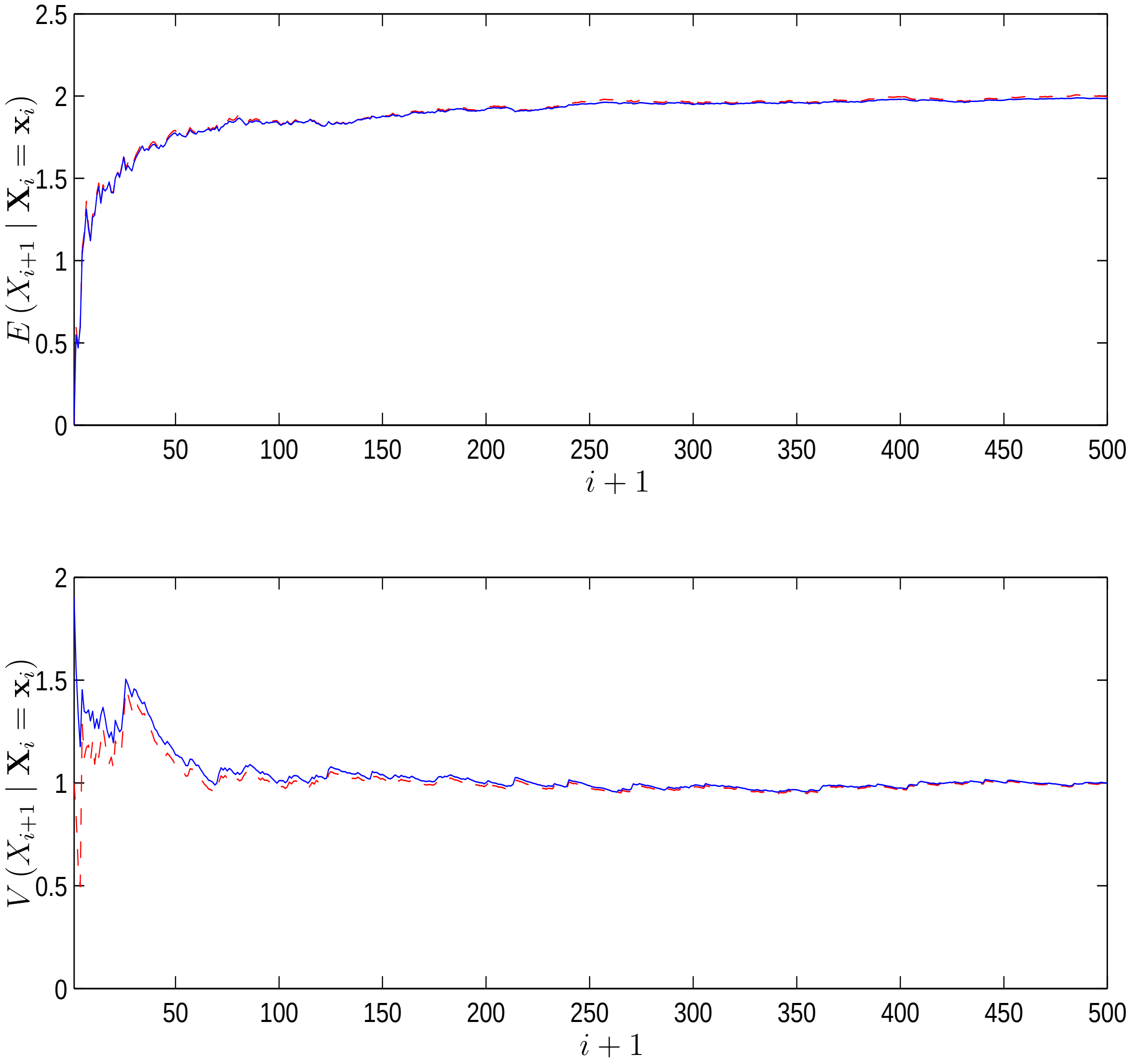

Figure 11: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{500} \sim N_{dig}(2, 1)$.

the digital computation. As the precision increases relative to μ, we expect to see more of a difference between the two conditional variances. Since we expect the range of possible X values specified by conjugate theory to be beyond the range in which $\mathbf{R}(X)$ is fully specified.

3.4.7 Case 2F: Observations Generated from $\mathbf{X}_N \sim N_{dig}(2, 0.25)$

Figure 13 shows the results of a typical case when the generating values of μ and π are 2 and 0.25 respectively. The number of observations drawn is 5000. The two

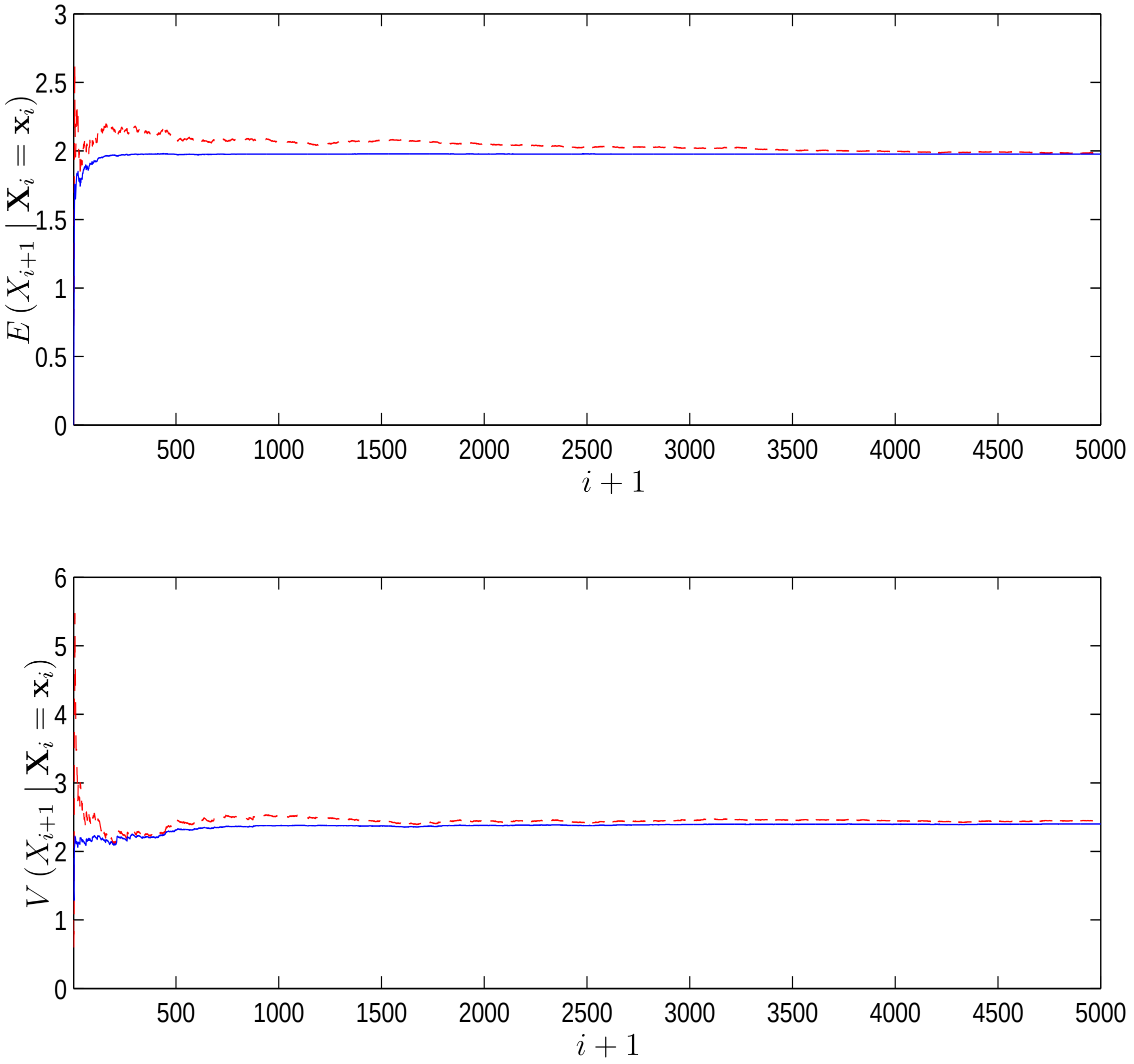

Figure 12: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{5000} \sim N_{dig}(2, 0.4)$.

conditional expectations are both similar, and are approximately equal to 1.9. In this case the mean of the conditional distribution of observations is 2, but both the computed and approximated conditional values of $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ are less than 2. This is because the range of $\mathbf{R}(X)$ means it is unlikely there will be any observations greater than 6, in fact $P(X > 6) = 0.0009$ (4dp). Whereas we are comparatively more likely to observe a value of X_i less than -2 since $P(X < -2) = 0.0243$ (4dp).

The approximation of the conditional variance is larger than the variance attained through digital computations, and both are less than 4. The computed conditional variance is approximately 2.5, as expected considering the maximum

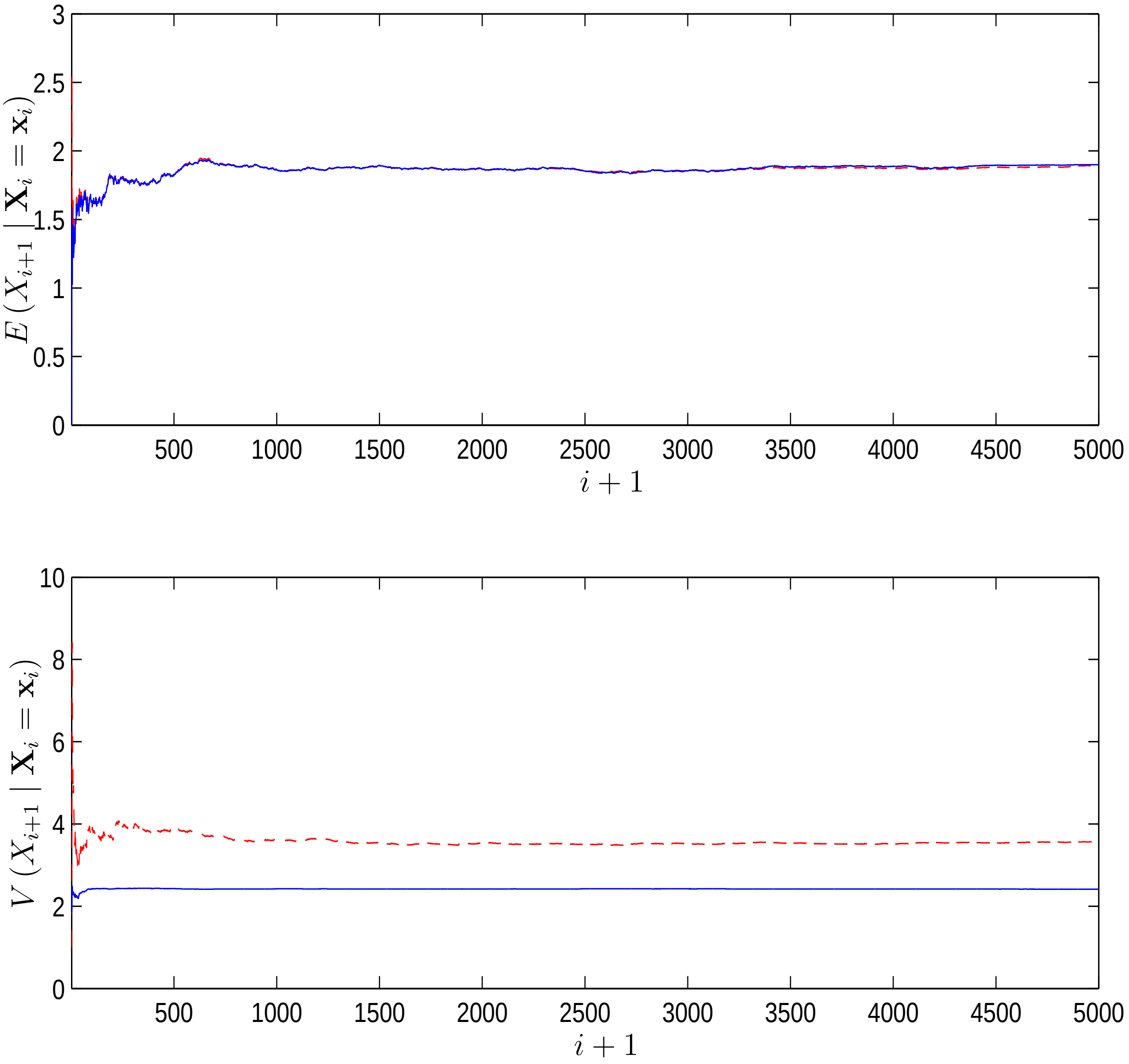

Figure 13: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{5000} \sim N_{dig}(2, 0.25)$.

value of $\mathbf{R}(\mu)$ is 2.5. The conjugate approximation gives a conditional variance of 3.6, significantly less than the approximation of 3.9 attained in Case 1C. The two components of the conjugate approximation of $V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ whose value will alter from Case 2C to Case 2F are s_i^2 and $\bar{x}_i$, both of these components will have lower values in Case 2F than Case 2C. Thus is because, although $\pi = 0.25$ in both cases, the number of distinct elements of $\mathbf{R}(X)$ that we could reasonably expect to observe is smaller in this case, as discussed in the paragraph above.

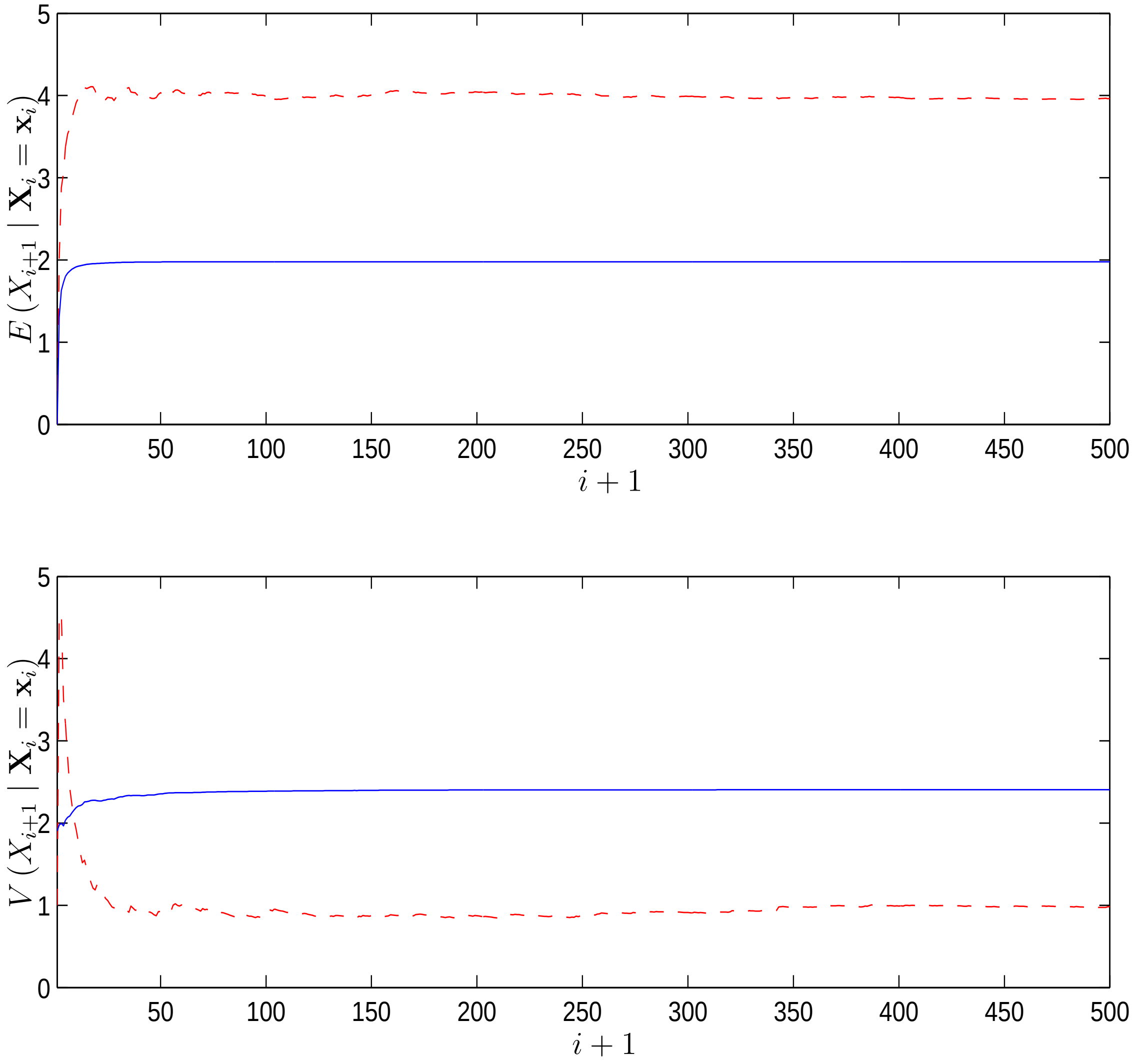

Figure 14: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{500} \sim N_{dig}(4, 1)$.

3.4.8 Case 2G: Observations Generated from $\mathbf{X}_N \sim N_{dig}(4, 1)$

Figure 14 shows the result of a typical example when 500 observations were generated from a digitised Normal mass function with parameter values of $\mu = 4$ and $\pi = 1$. The generating value of μ is not an element of $\mathbf{R}(\mu)$. Thus, if a researcher specified the prior mixture mass function considered in this Section, they would be surprised by the location of this set of observations.

The values of the conditional expectations are unsurprising. The two expectations have both reached stability more quickly than in either Case 2A or Case 2D.

$E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ for the conjugate approximation is just below 4, the mean of the conditional distribution of X values. The computed expectation is practically 2, the maximum value of $\mathbf{R}(\mu)$.

An interesting point to note is that in Case 1C, where we dealt with a one-stage prior and $\mathbf{R}(X)$ was fully specified on $[-4, 4]$, observations were also selected from a $N_{dig}(4, 1)$ mass function. The conjugate approximation was approximately 3.3. Now $\mathbf{R}(X)$ is fully specified on $[-6, 6]$, thus the observations drawn will have larger variance. If we were to rerun Case 2G with $\mathbf{R}(X)$ fully specified on $[-4, 4]$, we would obtain the conjugate approximation $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) = 3.32$.

The conditional variance is much larger than it was in Case 2D, when $\mathbf{X}_N \sim N_{dig}(2, 1)$. To see how $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ changes as μ increases, $V(X_{10000} \mid \mathbf{X}_{9999} = \mathbf{x}_{9999})$ was calculated for cases when observations were generated for selected digitised $N(\mu, 1)$ distributions. Results are listed in Table 1.

The approximated conditional variance decreases increasingly rapidly as μ increases. This is expected because as μ increases the possibility that an observation value could come from part of $\mathbf{R}(X)$ that is not fully specified increases. If we were to plot a bar graph of observations drawn from $N_{dig}(2, 1)$ they would appear as if they are drawn from a Normal mass function, but if we were to plot the observations from $N_{dig}(4, 1)$, the bar graph would be clearly truncated. This idea was investigated as part of Section 2.4.1, see Figure 1 for illustration.

In contrast the digitally computed conditional variance increases as μ increases. $V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right)$ is close to one when $\mu = 2$. As μ increases from 2.3 to 3.3 the increase is particularly rapid. When $\mu = 3.3$ the conditional variance has become as large as the elements of $\mathbf{R}(\mu)$ and $\mathbf{R}(\pi)$ will allow it to.

3.4.9 Case 2H: Observations Generated from $\mathbf{X}_N \sim N_{dig}(4, 0.4)$

Case 2H, where 5000 observations are generated from $N_{dig}(4, 0.4)$, produces similar results to Case 2G. The computed conditional expectation is still approximately 2, but the approximate expectation has dropped to 3.75, reflecting the non-symmetrical nature of the observations about 4 due to the smaller precision, see Section 2.4.5 for comments and Figure 16 for illustration.

The lower panel of Figure 15 demonstrates that the computed conditional vari-

μ	Digital Computation of Conditional Variance	Conjugate Approximation of Conditional Variance
2.0	1.0553	1.0269
2.1	1.0592	1.0271
2.2	1.0593	1.0264
2.3	1.0721	1.0044
2.4	1.1395	1.0039
2.5	1.2296	1.0035
2.6	1.3743	1.0018
2.7	1.5300	1.0240
2.8	1.6200	1.0024
2.9	1.7901	1.0008
3.0	1.9743	0.9972
3.1	2.1567	0.9944
3.2	2.3486	0.9895
3.3	2.4052	0.9855
3.4	2.4052	0.9789
3.5	2.4052	0.9682
3.6	2.4052	0.9616
3.7	2.4052	0.9495
3.8	2.4052	0.9366
3.9	2.4052	0.9057
4.0	2.4052	0.8910

Table 1: Digital computations and conjugate approximations of $V(X_{10000} \mid \mathbf{X}_{9999} = \mathbf{x}_{9999})$ when $X_{9999} \sim N_{dig}(\mu, 1)$.

ance is slightly smaller than 2.5, the maximum value allowable by $\mathbf{R}(\pi)$. The approximate variance has dropped to approximately 1.9.

3.4.10 Case 2I: Observations Generated from $\mathbf{X}_N \sim N_{dig}(4, 0.25)$

The conditional expectation and variance of a typical example when 5000 observations are generated from $N_{dig}(4, 0.25)$ are similar to results from Case 2H, as shown in Figure 16. A researcher who had specified the prior mass functions, $f(\mu \mid \pi) \sim N(0, \pi)$ and $f(\pi) \sim \Gamma(2, 2)$, would be very surprised to observe this data. For the conjugate conditional approximation $E\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) \approx 3.5$ and $V\left(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i\right) \approx 2.8$. Neither of these values are surprising considering the

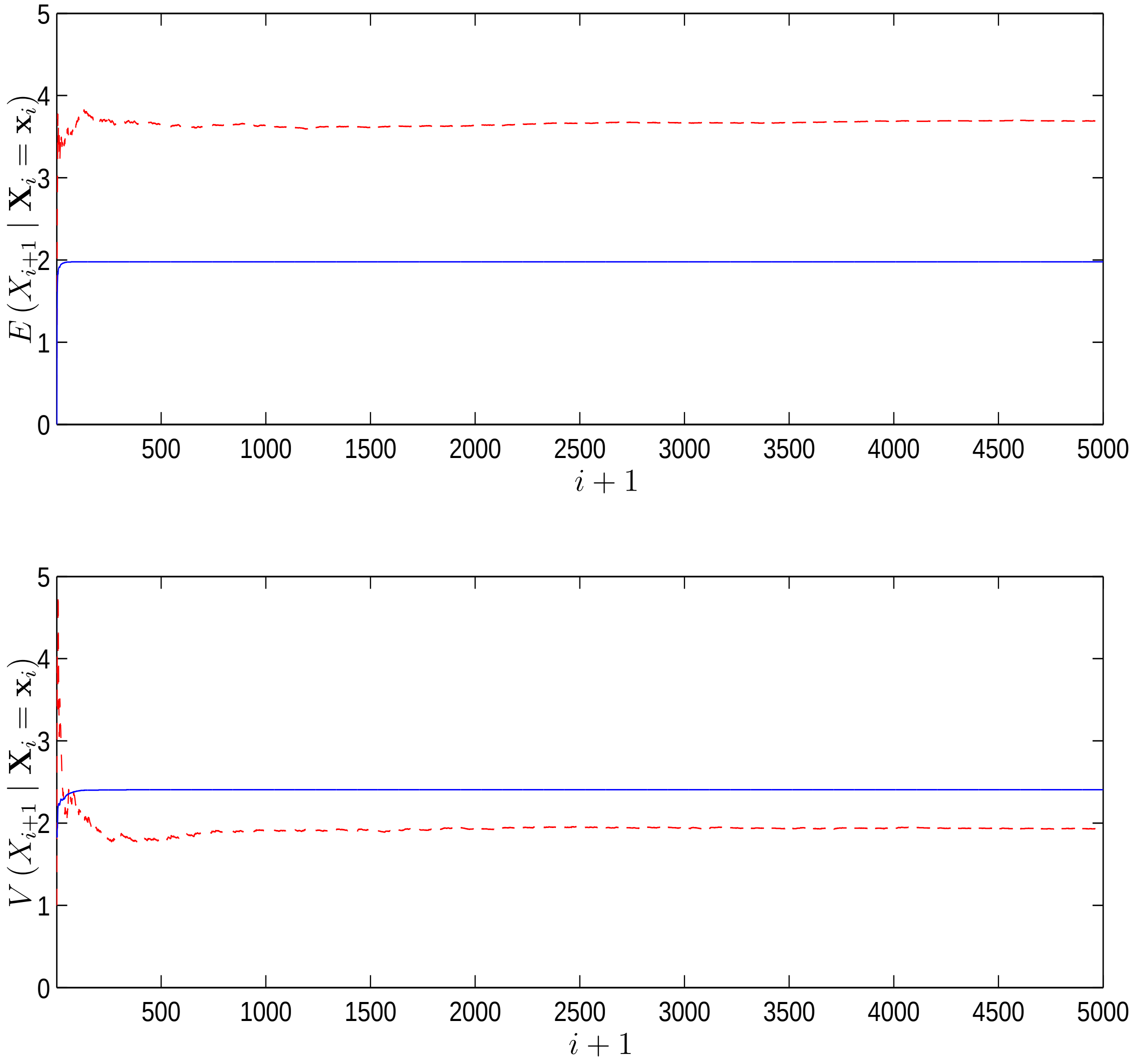

Figure 15: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{5000} \sim N_{dig}(4, 0.4)$.

range of observations values that are likely to be observed.

4 Summary

In this Report we have investigated how well continuous conjugate theory can approximate real discrete mass functions in various measurement settings We have described a procedure for assessing the value of conjugate continuous approximations in real problems where mixture digital mass functions can be specified.

Well known continuous distributions were digitised, and the means and variances

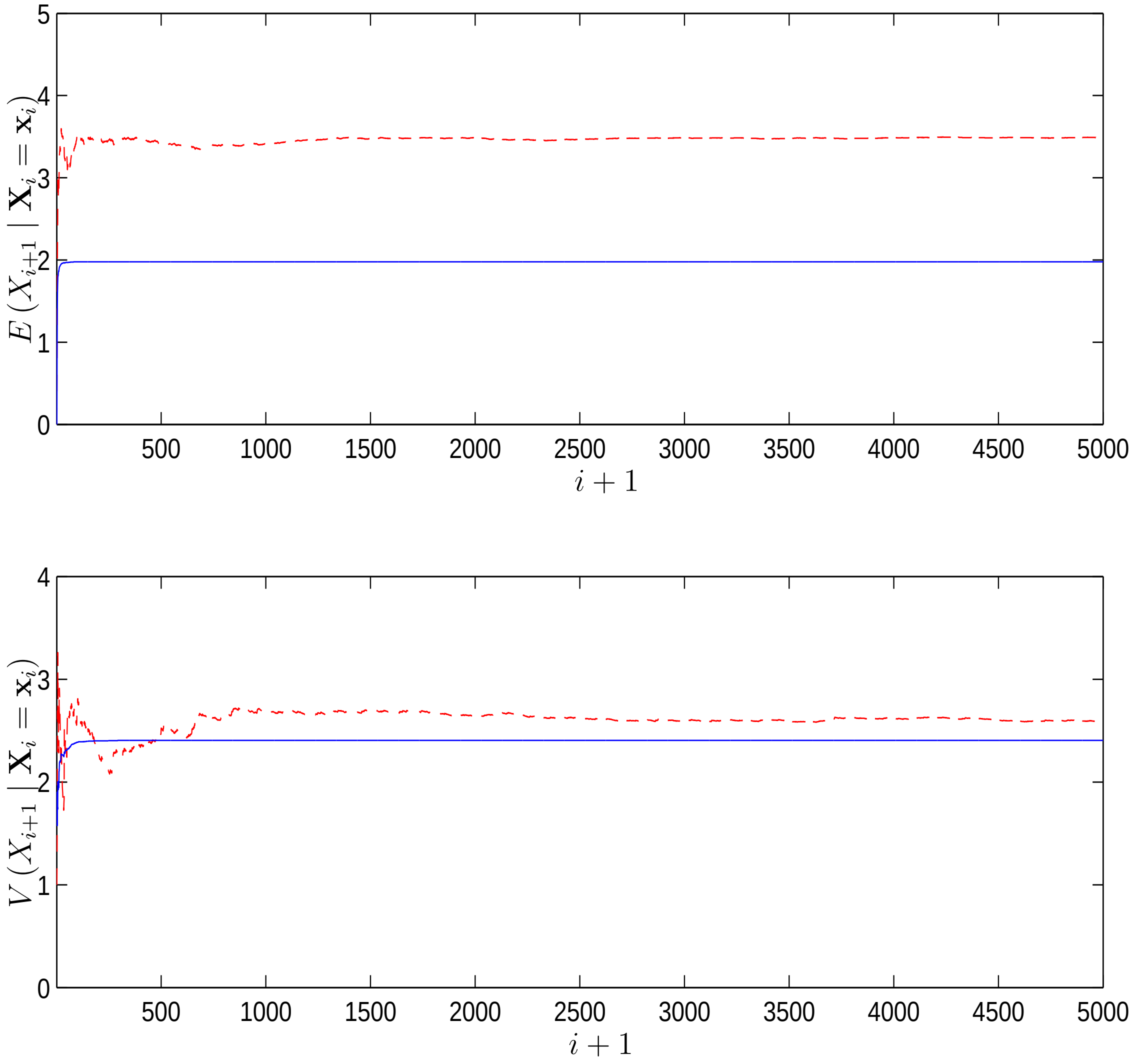

Figure 16: Conditional Expectations and Variances of the Digital Mass Function (solid blue line) and Conventional Conjugate Theory (dashed red line) when $\mathbf{X}_{5000} \sim N_{dig}(4, 0.25)$.

of their posterior mass functions computed. Conventional conjugate theory was used to approximate the means and variances of posterior densities. Our interest has centred on how well digital Normal mass functions and digital parametric mixtures are approximated by continuous mixture-Normal and Normal-Gamma mixture-Normal distributions for such items as $E(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$ and $V(X_{i+1} \mid \mathbf{X}_i = \mathbf{x}_i)$.

We have observed that when the researcher records observation values that are similar to what they have expected to observe, the discrete digital calculations and continuous approximations are very similar. That is, if the mode of the observations is close to the mode of the prior mixing mass function, and if the variance is small relative to the range of $\mathbf{R}(X)$, the actual calculations and the approximations are almost indistinguishable. When a researcher records observations they would be surprised by, for example, the cases when we defined the generating value of μ to be close to the extremities of $\mathbf{R}(\mu)$, the continuous approximations are larger (absolutely) than the digital calculations. When observations are such that the researcher is really surprised, the approximated conditional moments do not work at all well.

Acknowledgements

This work was undertaken as part of Robert Ware's doctoral thesis. He was partially supported by the University of Canterbury Keith Laugesen Scholarship and a Mathematics and Statistics Departmental Scholarship. Thank you to Doris Barnard and Alistair Smith for many helpful comments.

References

Lad, F. (1996). *Operational Subjective Statistical Methods. A Mathematical, Philosophical and Historical Introduction.* Wiley-Interscience.

Ware, R. and Lad, F. (2003). Flood Frequency Analysis of the Waimakariri River. Technical Report UCDMS2003/17, Department of Mathematics and Statistics, Univeristy of Canterbury, Christchurch, N.Z.