Defining Pseudo-Residuals for Multivariate HMMs

Tazman Libson

2024-04-5

Introduction

Throughout the world there are many modelling problems where observed data is impacted by unobserved factors. A person's mood impacts their decisions. An animal's hunger impacts its movement. Unknown processes impact the frequency of earthquakes. A way to account for this impact is representing this influence by a list of states which determine the observed variable. This kind of model is called a hidden Markov model.

Hidden Markov models (HMMs) are a class of stochastic models with a wide range of applications in ecology (Conners 2021), voice recognition (Boruah and Basishtha 2013), finance (Oelschläger and Adam 2023), and many other fields. They are most often used with either temporal or spacial data. There are two main components for a hidden Markov model, an unobserved process, and an observed process/variable. The unobserved process is a discrete set of states, described by a Markov chain. The interpretation of these states depend on the application of the model. For example, in finance, an interpretation for a 3 state model would be a "bearish"/growth state, a "bullish"/decay state, and a mixed middle state (Oelschläger, Adam, and Michels 2024). The key feature of these states is that they cannot be directly observed. For example, there is not a way to say with certainty which of the mentioned states the market is in at any given time. For each state of the unobserved process there is a state-dependent distribution. These distributions are what determine the values of the observed variable. Again in a financial context, an observed variable could be the returns of a stock. This report uses financial data as a case study, however the methods described can be used for any application of HMMs.

There are many decisions one has to make when fitting HMMs. One has to decide on the number of states and the state dependent distributions before doing any model fitting. Thus, model assessment is very important. There are some broader model selection methods like information criterion (e.g AIC, BIC) one can use to inform this decision. While these are useful, there are additional model evaluation methods that we can use. There is a model diagnostic for HMMs called pseudo-residuals. This method is well defined (Walter Zucchini 2016) and is commonly used in the application of HMMs. However, is only described for univariate data (i.e. data where there is only one observed variable per time). This report is going to extend pseudo-residuals for multivariate data (i.e. data where there is more than one observed variable per time).

In section 1, HMMs are explicitly defined and the required parameters for each model are identified. In section 2, the model fitting process using `nlm` is described and a novel reparameterization is described for multivariate-normal HMMs. In section 3, the definition for pseudo-residuals is presented. Then this definition is extended for multivariate data using 2 methods which will be called vector and element pseudo-residuals. Section 4 describes the results of the case study and the behaviour of the vector and element pseudo-residuals for the fitted model of the case study. Section 5 describes the results of an experiment using simulated data to further investigate the properties of vector and element pseudo-residuals. The code that was written for sections 2-5 can be found in the github repository linked at the end of the paper.

Case Study: Returns for 4 Tech Companies.

For this report, daily returns data for 4 companies: Apple, Microsoft, Meta, and Intel between Feb 1st 2019 and Feb 1st 2024 were used (*NASDAQ* 2024). The returns of each stock are going to be modelled.

The returns are defined as $100\log(s_t/s_{t-1})$, where s_t is the price on day t (Walter Zucchini 2016). The Hidden Markov models fitted will be fitting multivariate data. In other words, the returns for all 4 stocks are being modelled simultaneously. This will allow the model to take into account the correlations between stocks. As seen in the small slice of the returns for all 4 stocks are plotted in Figure 1, the return values rise and fall together at some times, at other times they diverge. In financial modelling, stocks are often correlated, and have changing correlation over time (Tobias Preis 2012). Multivariate hidden markov models will accommodate for this. Let us now move onto defining HMMs.

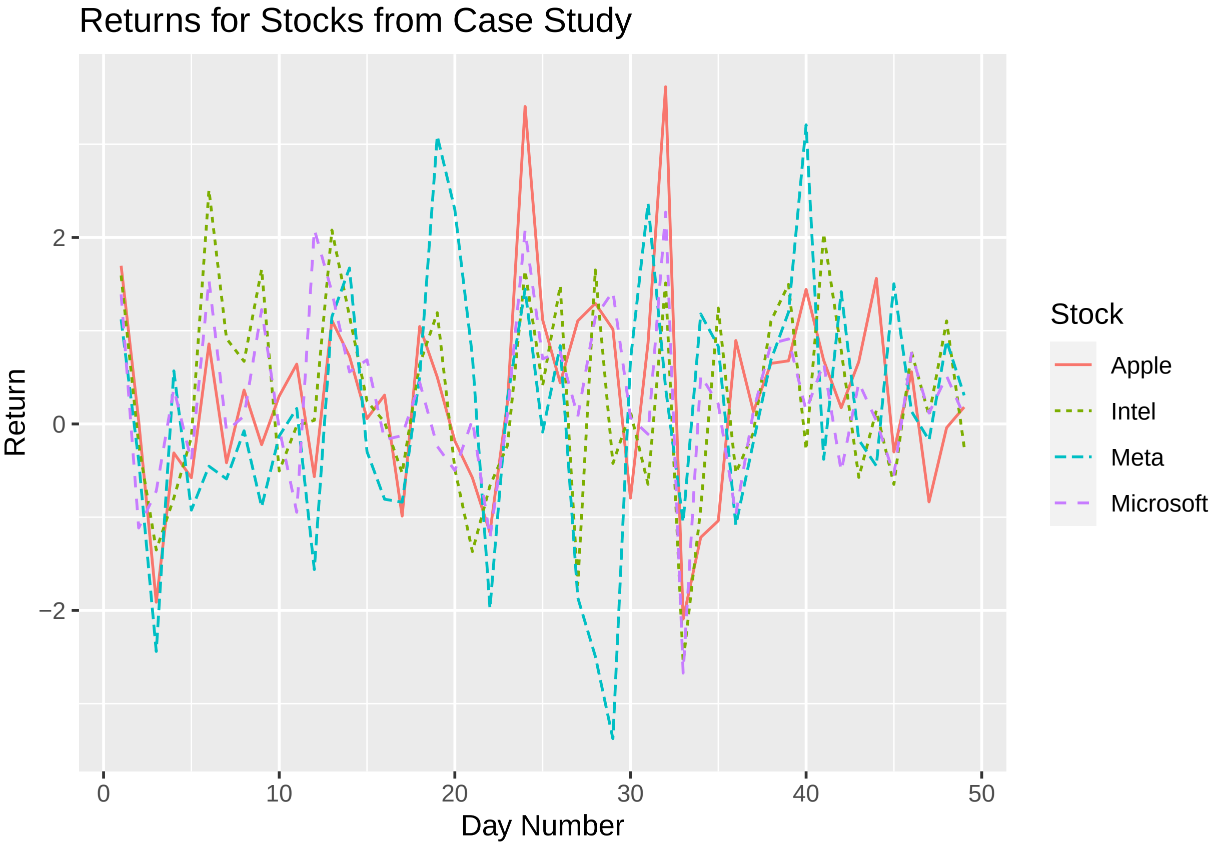


Figure 1: Log returns for Apple and Microsoft between Feb 5th 2019 and May 6 2019. There are times when the returns rise and fall with eachother, and others where they diverge. This visually demonstrates the changing correlations of the return values. These changing correlations can be accounted for by HMMs.

Section 1: Model Definition

The following description of Markov chains and their relation to the state dependent distribution comes from a widely cited book on HMMs (Walter Zucchini 2016). Each HMM has an unobserved process and an observed random variable. The unobserved process affects the distribution of the observed random variables. First the unobserved process will be described followed by the unobserved process.

Unobserved Markov Chain

The unobserved states and the transitions between states for HMMs is described by a Markov chain. A Markov chain is a sequence of discrete random variables where each time has a state from a discrete set of possible states. The probability for a given time to be in any state is only affected by the state of the previous time. This chain will have m states, so the possible states are $1, 2, ..., m$. The state at time t will be written as c_t. The probability from going from state i at time t to state j at time $t+1$ will be referred to as p_{ij}. These transition probabilities are used to make the one step transition probability matrix, $\mathbf{\Gamma}$ which will be used to calculate how the state of the Markov chain evolves over time.

$$\mathbf{\Gamma} = [p_{ij}]; \quad p_{ij} = \mathbb{P}[c_t = j | c_{t-1} = i]$$

For all our cases, the markov chains will all be homogeneous, meaning that $\mathbf{\Gamma}$ is independent of time. This is a common assumption in many applied uses of HMMs. It to simplifies the model and is required for the likelihood calculation, and thus the entire model fitting process.

The transition probability matrix will allow for the expected evolution of states to be calculated given an initial state distribution. The assumption of time homogeneity makes this process much simpler. A state distribution $\mathbf{u}_t$ is a vector of length m, the number of states, where u_i is the probability at time t for $c_t = i$.

$$\mathbf{u}_t = \{u_1, u_2, ..., u_m\}, \quad u_i = \mathbb{P}[c_t = i]$$

To get the state distribution for the following time, one takes the product of the current state distribution and the transition probability matrix.

$$\mathbf{u}_t \mathbf{\Gamma} = \mathbf{u}_{t+1}$$

The stationary distribution of a markov chain is a state distribution where, when multiplied with the transition probability matrix, remains unchanged. It is defined as follows:

$$\mathbf{u}\mathbf{\Gamma} = \mathbf{u}$$

For model specification, one has to either indicate a specific initial distribution, or indicate that the model is stationary, and the stationary distribution will be used as the initial distribution. It can be shown that every markov chain that will be used in our models has a stationary distribution(Privault 2013). Since every Markov chain used for our models has a stationary distribution, assuming the Markov chain is stationary will work for every model instead of specifying an initial distribution. For the case study, it was assumed that the model starts at its stationary distribution.

State Dependent Distributions for the Observered Variable

For all the models in this paper, the state dependent distributions will be an n-dimensional multivariate normal. This is a common distribution to use in financial contexts (Walter Zucchini 2016). For every state $i \in (1, ..., m)$, the state-dependent distribution, $p_i(\mathbf{x}_t)$ will be defined as follows:

$$p_i(\mathbf{x}_t) = \mathbb{P}[\mathbf{X}_t = \mathbf{x}_t | c_t = i] \qquad \mathbf{X}_t | c_t = i \sim N(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$$

For every state $i \in \{1, .., m\}$ the probability density function will have a $n \times n$ variance-covariance matrix $\mathbf{\Sigma}_i$ and a vector of means of length n μ_i. Since this distribution is multivariate, the input x_t is a vector of length n.

To summarise, for an m state HMM with n-dimensional multivariate normal state dependent distributions will need to define the following values:
* Transition Probability Matrix, $\mathbf{\Gamma}$. An $m \times m$ matrix
* Initial Distribution/Stationary Distribution, $\boldsymbol{\delta}$, a vector of length m
* Means, $\boldsymbol{\mu}_i$ a vector of length n for each state $i \in \{1, ..., m\}$
* Variance-Covariance Matrix,$\mathbf{\Sigma}_i$, a $n \times n$ matrix for each state $i \in \{1, ..., m\}$

Now let us go through how one can find values for these parameters.

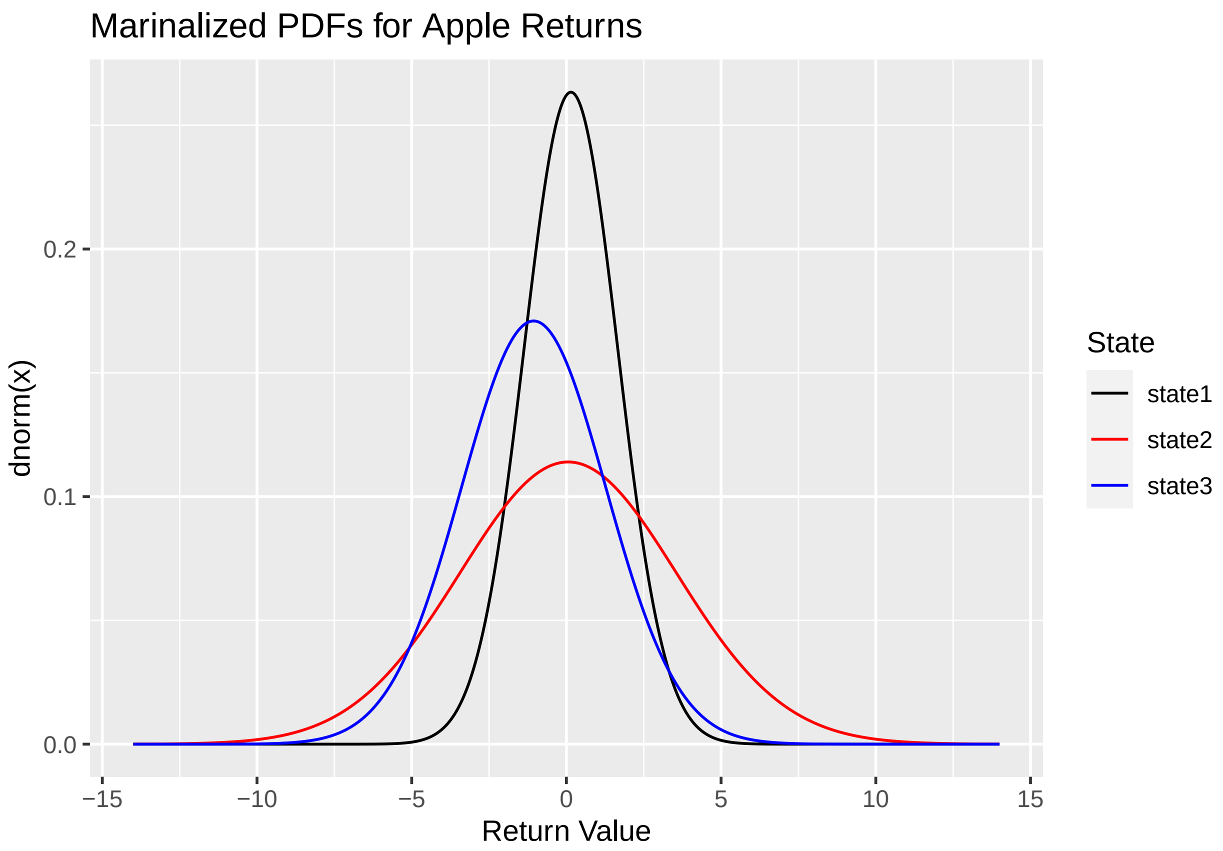


Figure 2: Marginalized State Dependent Distributions for the Apple stock from the fitted model. Although each state has a normal distribution, each one has a different mean and variance. For the technology stocks that were fitted using the model, each state has a different 4-dimensional normal distribution. See appendix A for specific model parameter values for the case study.

Section 2: Model Fitting

Model fitting was done by maximum likelihood estimation, where the likelihood was calculated for an initial model, then using the non-linear maximization function `nlm`, one finds the parameter values which finds the model where observing the data has the largest likelihood.

Likelihood Calculation

It can be shown that the likelihood of a series of observations, L_T from a HMM can be calculated recursively through matrix multiplication (Walter Zucchini 2016). Here $\mathbf{X}^{(T)}$ indicates all the observations from $t = 1, ..., T$. Since each observation has n elements, $\mathbf{X}^{(T)}$ is a $n \times T$ matrix of observations. $\mathbf{x}_t$ indicates the length n vector of observations at time t.

$$L_T = \mathbb{P}[\mathbf{X}^{(T)} = \mathbf{x}^{(T)}] = \boldsymbol{\delta} P(\mathbf{x}_1)\boldsymbol{\Gamma} P(\mathbf{x}_2)\boldsymbol{\Gamma} P(\mathbf{x}_3)...\boldsymbol{\Gamma} P(\mathbf{x}_T)\mathbf{1'}$$

Where $P(\mathbf{x}_i)$ is a $m \times m$ diagonal matrix of the state dependent distributions $p_i(\mathbf{x}_t)$, and $\mathbf{1'}$ is a column vector of 1s of length m.

The calculation takes into account the probability of each observation under each state-based distribution and is scaled by the probability of being in each state at the time of each observation. To avoid underflow, the log-likelihood is calculated for the model fitting functions.

Reparameterization

The optimization function used for likelihood maximization, `nlm`, needs unrestricted parameters (R Core Team 2023a). Unrestricted here means the function that is being optimized needs to use parameters that can take any real number. The means of the state dependent distributions, $\boldsymbol{\mu}_i$, are unrestricted. The other parameters, the variance-covariance matrices $\boldsymbol{\Sigma}_i$ and the transition probability matrix $\boldsymbol{\Gamma}$, all have several restrictions. These parameters will be have to be transformed. In the following descriptions, “working” parameters will be the reparameterized values, “natural” parameters will be the non transformed value.

Transition Probability Matrix The following method of reparameterization for the transition probability matrix has been previously described (Walter Zucchini 2016). The values of the transition probability matrix are all between 0 and 1 inclusive. The rows also must all sum to 1. So we can set the diagonal elements to be 1 minus the sum of the other elements of their row, meaning the non diagonal elements are the ones being estimated. In total, there are only $m(m-1)$ free parameters for the transition probability matrix. The natural parameters are first transformed to be on the non-negative reals by dividing them by the diagonal values. These values are then mapped to the entire real line by the taking the log function, which maps the positive reals onto the entire real line. Thus they are unrestricted. Here the natural transition probabilities will be p_{ij} and the working transition probabilities will be τ_{ij}.

$$\tau_{ij} = \log(\frac{p_{ij}}{p_{ii}}), \quad i \neq j; \; \tau_{ii} = 1 \; \forall \; i \in \{1, ..., m\}$$

$$p_{ij} = \frac{\exp(\tau_{ij})}{1 + \Sigma_{i \neq k} exp(\tau_{ik})}$$

Variance-Covariance Matrix The following reparameterization for the variance-covariance matrix is newly described by this paper. The restrictions on the variance-covariance matrices make it inconvenient to directly transform the natural covariances into working parameters. The matrices (remember there is a separate matrix for each state) are positive definite, meaning they are symmetric and their eigenvalues are positive. In addition, the square of all of the non-diagonal elements cannot be larger then the product of both of the variances of its row or column number (i.e. for covariance $[\text{Cov}(X_i, X_j)]^2 \leq \text{Var}(X_i)Var(X_j)$). This comes from the Cauchy-Schwartz inequality (Silvey 1975). Thus every non diagonal element has a unique range of possible values depending on the values of variances for each of the variables of the multivariate-normal.

Instead of finding the variance-covariance matrices directly, one can estimate the variances and correlation matrices separately and then get the variance-covariance matrices from there. The correlation matrix is just a rescaled variance-covariance matrix. Estimating the variances and correlation matrices separately is equivalent to estimating the variance-covariance matrix.

The correlation matrices are also all positive definite. In addition, the values on the diagonal are all 1, since each variable (in the case study the return of each stock) always has a correlation of 1 with itself. The non-diagonal values all have values between -1 and 1 inclusive. Now each correlation value has the same range of possible values, as opposed to the unique intervals seen with covariances. The correlation matrices are symmetric so only the non-diagonal upper triangular values of the correlation matrix need to be estimated. A scaled tan function is used to map the correlation values onto the reals. Here the natural values will be k^h_{jh} and the working values will be κ^h_{ij} with ij indicating the position in the correlation matrix and i indicating the state for which the correlation is used.

$$\kappa^i_{jh} = \tan(\frac{\pi c^i_{jh}}{2}), \quad k^i_{jh} = \frac{2\arctan(\kappa^i_{jh})}{\pi}; h \neq j$$

The variances are all strictly positive so they can be reparameterized by taking the log of the natural parameters. The working variance will be σ^2_{ij}, the natural variance will be s^2_{ij}. Here i is one of the m states and j is one of the n observed variables.

$$\sigma^2_{ij} = \log(s^2_{ij}), \quad s^2_{ij} = \exp(\sigma^2_{ij})$$

Section 3: Pseudo-Residuals

Pseudo-residuals are a commonly used model diagnostic for HMMs. The pseudo-residuals described by Zucchini et al are widely used by many fields which use HMMs (Conners 2021; Oelschläger and Adam 2023; Fernández-Fontelo 2019; Lintern 2023; McClintock 2024). Their two main uses are evaluating model fit and identifying outliers.

In the same way that residuals measure the deviation of observations to a model, pseudo-residuals measure how likely the observations under a the distribution that the HMM predicts for that time. These distributions are called conditional distributions. Each time t will have a different conditional distribution.

Conditional Distributions

In order to describe pseudo-residuals, we must define conditional distributions. First we need to define a small bit of notation. The observation vector (or matrix in the case of a multivariate model) $\mathbf{X}^{(-t)}$ will be defined as all of the observations from time $1, ...T$ except t. So it represents $\{X_1, X_2, ...X_{t-1}, X_{t+1}, X_{t+2}, ...X_T\}$. With that conditional distributions are defined as the following probability distribution (Walter Zucchini 2016):

$$f_t(\mathbf{x}) = \mathbb{P}[\mathbf{X}_t = \mathbf{x}|\mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}]$$

So conditional distribution at time t is the probability distribution of $\mathbf{X}_t|\mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}$ or the probability of observing a specific value under the model at time t given every other observation besides the observation at time t. These distributions can act as the model's predicted distribution for each time. In Figures 4 and 5, one can see two different representations of some of the conditional distributions for the model of

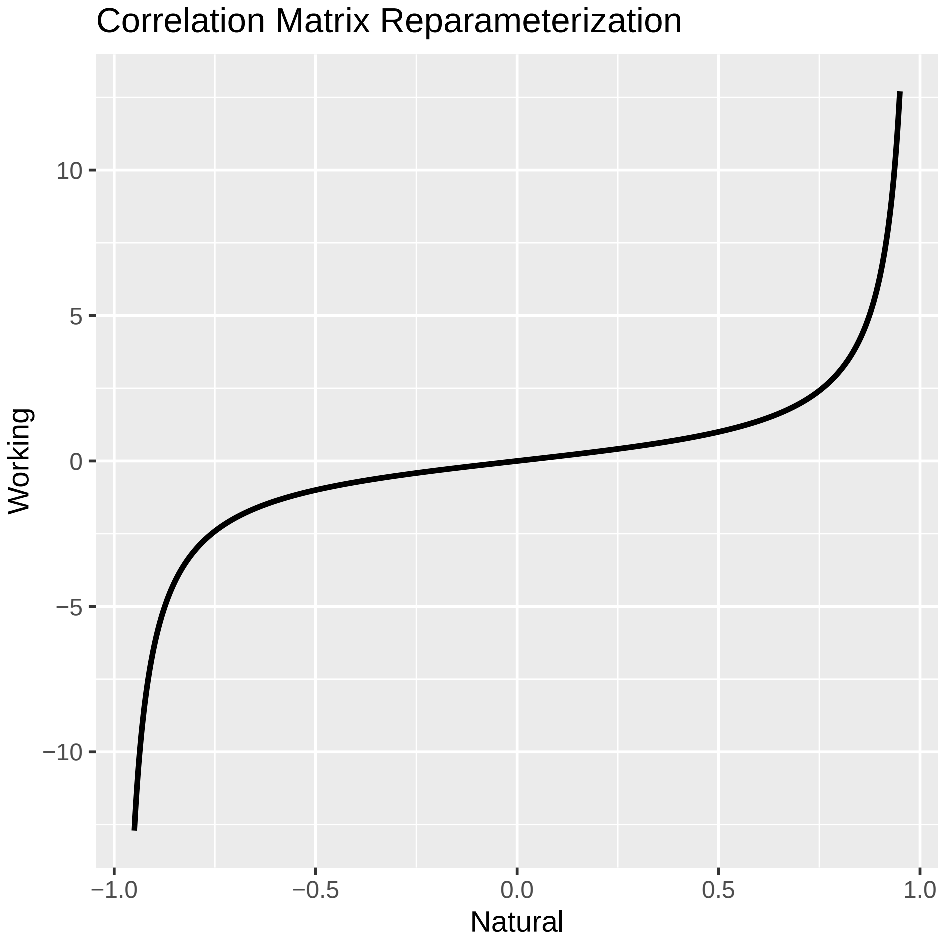


Figure 3: Graph of the natural correlations to working correlations. The correlation values can range from -1 to 1. The scaled tangent function maps these values onto the entire real line. Most natural correlation valuesare between -0.8 and 0.8.

the case study. Unfortunately the full conditional distributions for a given time cannot be easily displayed because of the dimensionality of the distribution (in our case the conditional distribution for each time is 4-dimensional). One can see that the conditional distributions change day by day.

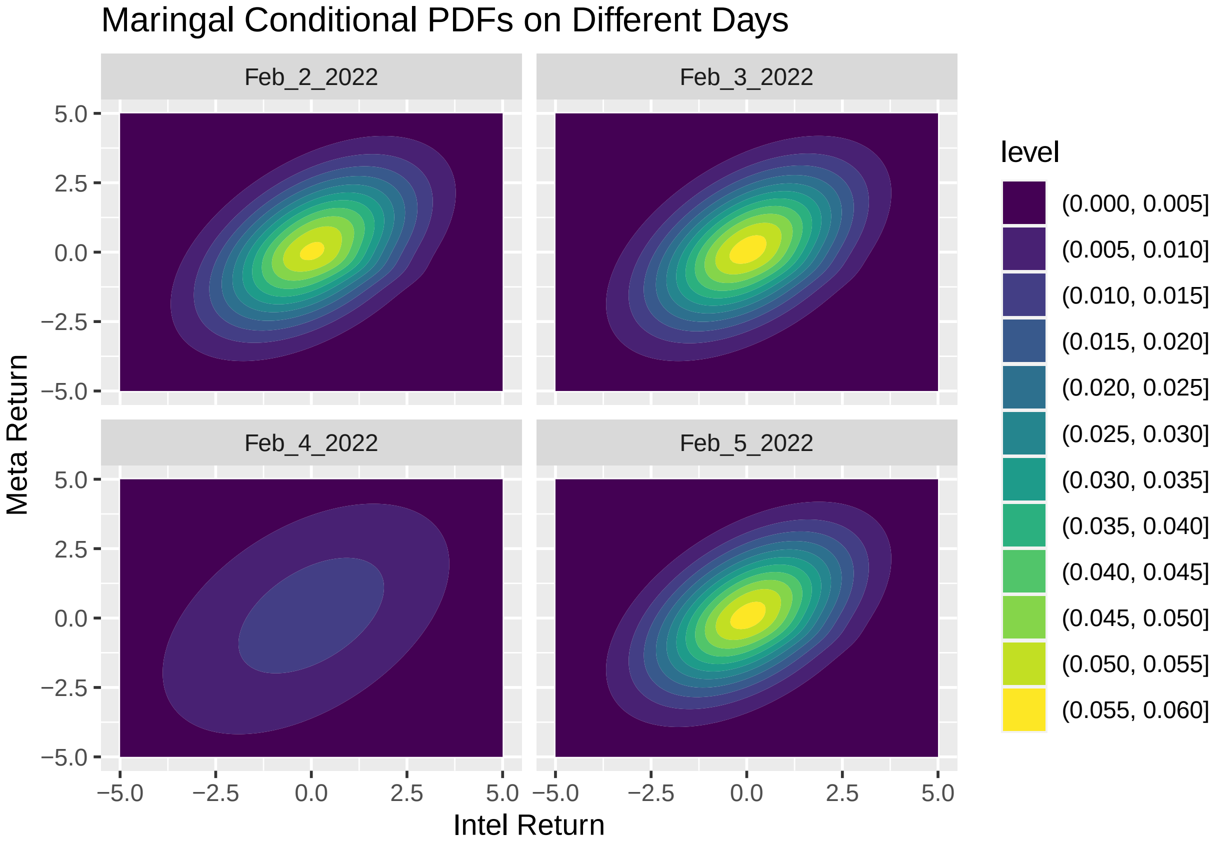


Figure 4: Marginalized Bivariate Conditional Probability Distribution Functions for Feb 2-5 2022. Each day has a different distribution. Here the colors indicating the color is the value of the conditional distribution at that point. On Feb 3, there is a massive drop in the Meta stock, which has a significant impact on the following day's distribution. The full conditional distributions for the case study cannot be displayed because they are 4-dimensional.

So far, only conditional probability density functions have been shown. For pseudo-residuals, the conditional cumulative density functions (conditional cdfs) will be needed. For now, the conditional cdfs will be only univariate. They will be extended later to be multivariate. Conditional cdfs will be indicated by $F_t(x)$ and will be defined as follows:

$$F_t(x) = \mathbb{P}[X_t \leq x | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}]$$

With conditional cdfs defined (see appendix B for further details), we can move on to the definition of pseudo-residuals. The definition for pseudo-residuals used for univariate data will be presented. Then this definition will be extended to multivariate data in two ways.

Uniform Pseudo-residuals

First we will define uniform pseudo-residuals. Uniform pseudo-residuals are the conditional cdfs for every time in the data set evaluated at every observation in the dataset. This will generate values that are between

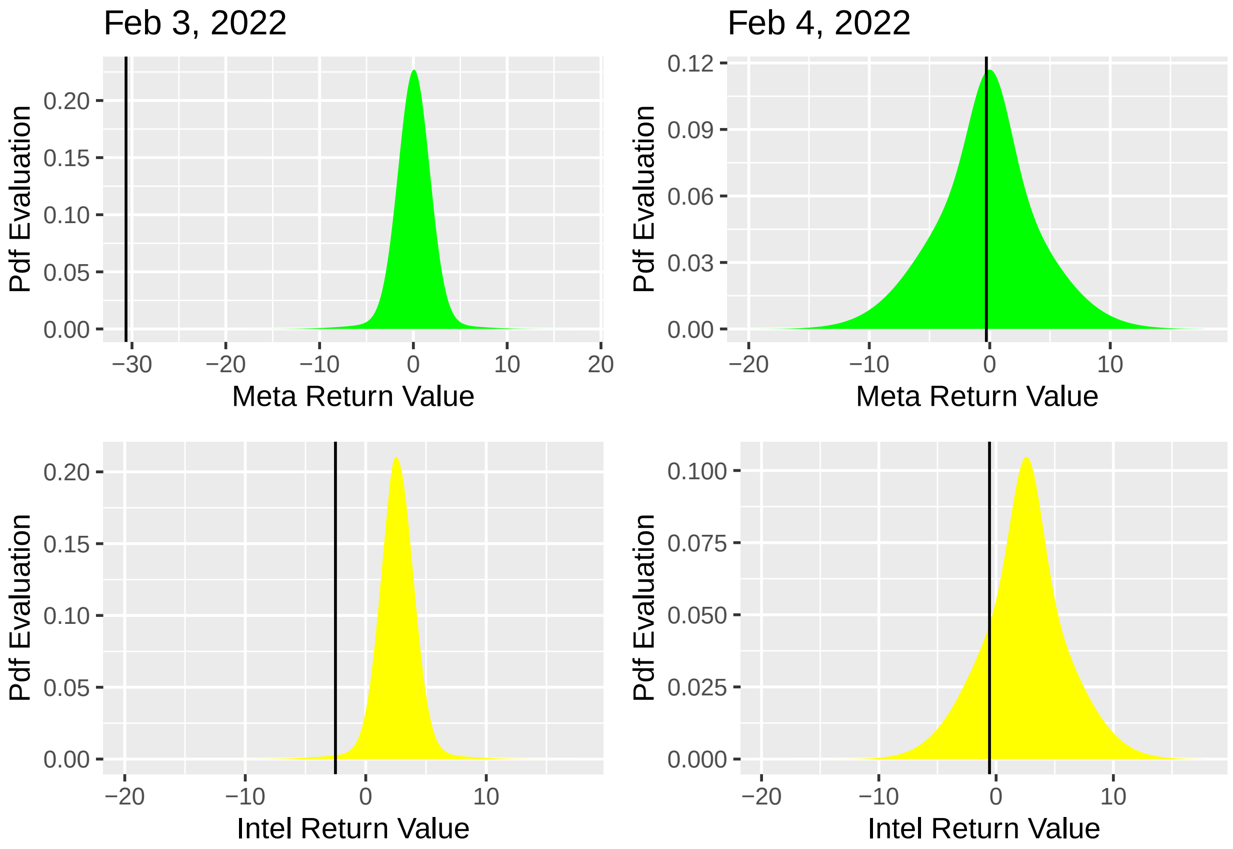


Figure 5: Marginalized conditional univariate conditional pdfs for Meta and Intel for Feb 3-4 2022. Once again can see the impact of theMeta stock drop where the variance of the univariate conditional pdfs both increase for the following day.

0 and 1, since they are the output of a cdf, and thus a probability. The utility here is that for a well fitted model, these uniform pseudo-residuals should be uniformly distributed.

$$F_t(X_t) \sim U(0,1)$$

There are some issues with uniform pseudo-residuals, mainly that they are not well suited for outlier detection. Outliers here would be defined as values on the fringe of their respective conditional distributions. For uniform pseudo-residuals, outliers would be observations with pseudo-residual values close to 0 or 1. Values that would be considered outliers, however, would be hard to distinguish from one another(e.g. distinguishing between 0.99 and 0.999)(Walter Zucchini 2016). To solve this issues we can transform uniform pseudo-residuals.

Normal Pseudo-residuals

With these uniform pseudo-residuals we can define normal pseudo-residuals. By taking the inverse standard normal function, Φ^{-1} (`qnorm` in `R`), of the uniform pseudo-residuals we get normal pseudo-residuals. With these pseudo-residuals, outliers are clearer to distinguish between one another. See Figure 6 for a visual demonstration of both kinds of pseudo-residuals for univariate data. Normal pseudo-residuals are standard normal distributed if the model fit is good.

$$\Phi^{-1}(F_t(X_t)) \sim N(0,1)$$

From now on we will only work with normal pseudo-residuals, so for convenience they will be referred to as only pseudo-residuals.

Pseudo-residuals for univariate data have been clearly established and have been widely used in the application of HMMs. Let us now proceed to defining pseudo-residuals for multivariate data. There is ambiguity on how to extend pseudo-residuals for multivariate data. The crux of the issue is how the data is inputted to the conditional cdf. Are we taking the pseudoresidual for a time, and thus using each variable simultaneously in the conditional cdf, or are we taking the pseudoresidual for each variable separately for each time. I have come up with two separate metrics to investigate this which will be defined as vector and element pseudo-residuals.

Vector pseudo-residuals

Vector pseudo-residuals are acquired by inputting the entire vector (hence the name) of observations for each time into the conditional cdf. Here $\mathbf{X}_t \leq \mathbf{x}_t$ will indicate $\{X_t^1, ..., X_t^n\} \leq \{x_t^1, ..., x_t^n\}$. In other words, the probability that each separate variable X_t^j is less than or equal to the observed value for that variable at that time x_t^j for all $j \in \{1, ..., n\}$.

$$F_t(\mathbf{X}_t) = \mathbb{P}[\mathbf{X}_t \leq \mathbf{x}_t | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}]$$

This will yield a single pseudo-residual value for each time, so there will be a total of T vector pseudo-residuals

Element pseudo-residuals

Element pseudo-residuals are acquired by inputting each element of the observation vector into a marginalized conditional cdf individually. When a variable is marginalized, its influence on the function is eliminated by taking the cdf over the entire possible range of values that the variable can take. This reduces the dimensionality of the distribution. For element pseudoresiduals, all but one of the variables are being marginalized which results in a separate 1 dimensional conditional cdf for each variable. These 1 dimensional conditional cdfs can be seen in Figure 5. In our case study, each stock is a separate variable, so each stock

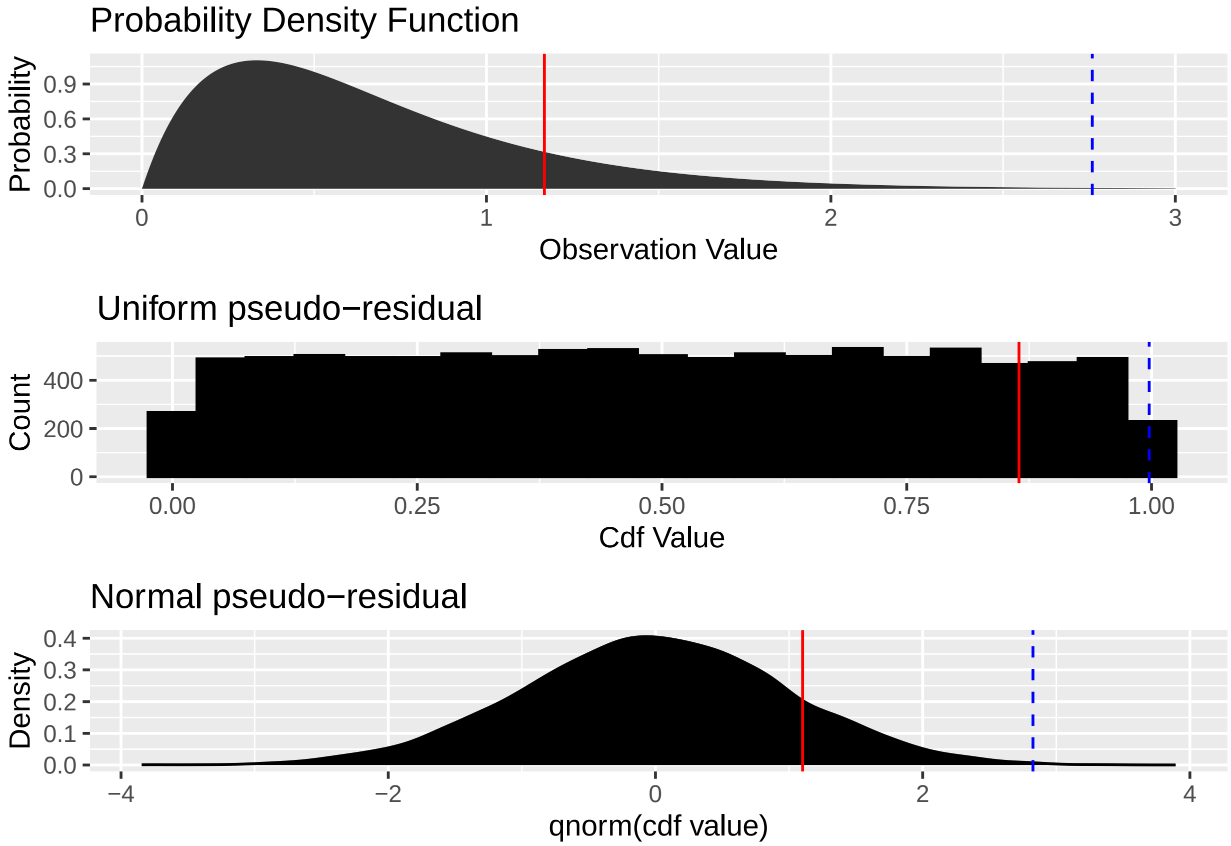


Figure 6: Demonstation of pseudo-residuals for a univariate conditional pdf (in this case just a gamma distribution). The solid and dashed lines represent observations as they are transformed into uniform pseudo-residuals by taking the conditional cdf of the values. This is expected to be uniformly distributed. Then the inverse normal is taken to get normal pseudo-residuals. Notice how visually, it is easier to evaluate normality than uniformity. Notice how the blue/dashed line and the red/solid line are quite far apart in the conditional distribution. The dashed line is significantly less likely than the solid line. In the uniform pseudo-residuals they don't have as much distance between them. For the normal pseudo-residuals, the two lines are more distinguishable

at each time would have a different marginalized conditional cdf. $X_t^j \in \mathbf{X}_t$ indicates an observation of one of the variables j at time t then the marginalized conditional cdf for X_j, F_t^j, will be defined as

$$F_t^j(X_t^j) = \mathbb{P}[X_t^j \leq x_t^j | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}]$$

This process will yield an element pseudo-residual for each observation of each variable. So there will be nT element pseudo-residuals.

Let us now examine how these two pseudo-residuals behave for the case study.

Section 4: Results for Case Study

The case study will be used to examine how the two kinds of pseudo-residuals behave as outlier detectors, and indicators for goodness of fit. First will be outlier detection.

Pseudo-residuals for Outlier Detection

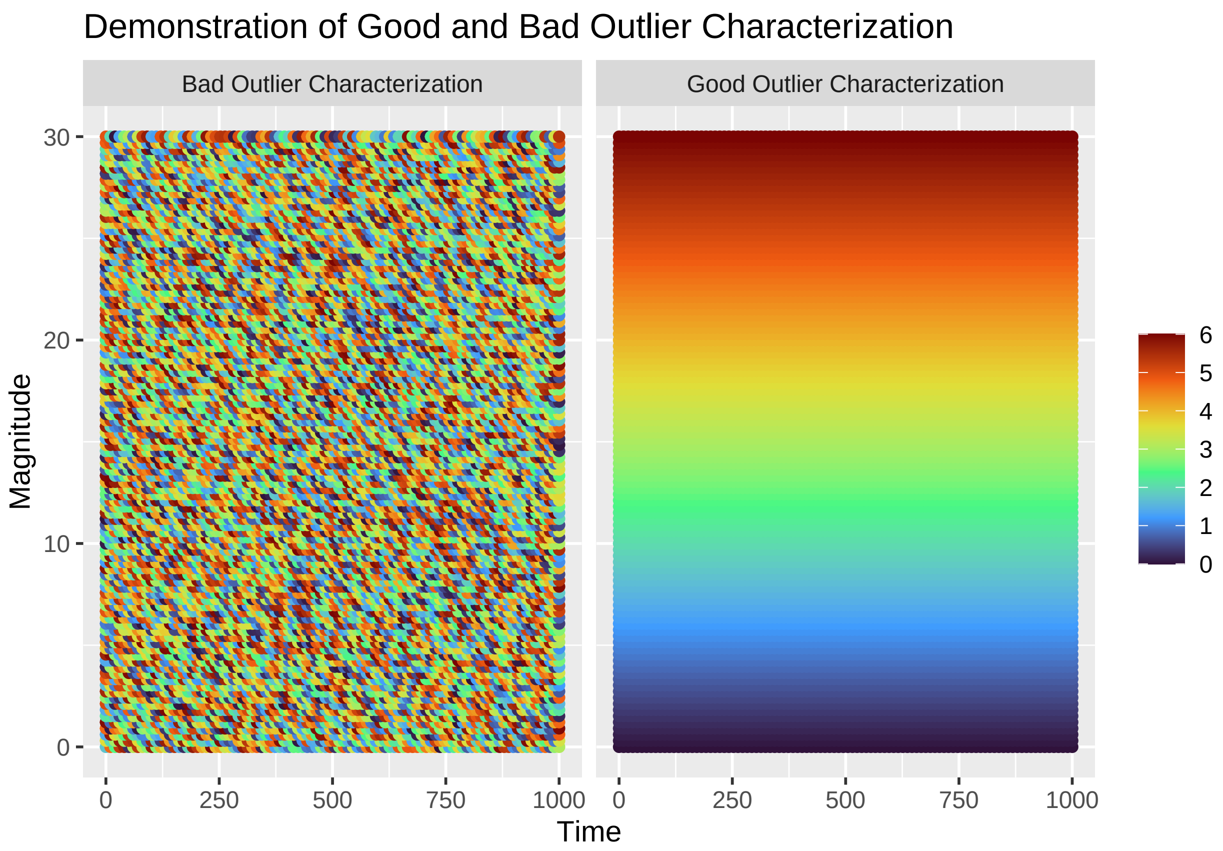
Outlier detection is important to modelling and has many uses (**Outlier1?**; **Outlier2?**). Having a metric for outlier detection helps with both model evaluation and also data analysis. For model evaluation, identifying outliers can be used to improve model fit and improve accuracy. Outliers can also provide insight into collected data. In our case study, there was already an idea of what an outlier was in the context of financial returns. For other uses of HMMs, for example animal movement, using pseudo-residuals for outlier identification could give some insights into what kinds of movement are uncommon in a dataset. Now let us move on to outlier detection for the case study.

For us to examine how element and vector pseudoresiduals identify outliers we must first state what we are viewing as outliers for our data. For the case study, outliers will be defined as returns with large magnitude. This comes from general intuition, large gains or losses in a single day for a stock aren't frequent and normally unexpected. This is also justified by the data. The mean returns for all 4 stocks over the 5 year period are relatively close to zero, so large magnitude returns are going to be several standard deviations away from the mean (see table 1).

Table 1: Summary Statistics for Returns Data

Stock	Mean	Var
Apple	0.1172	3.9994
Intel	-0.0101	6.1147
Meta	0.0674	7.6574
Microsoft	0.1066	3.6486

So large magnitude returns are outliers will be considered. For both types of pseudo-residual, a large magnitude pseudo-residual indicates an unlikely observation according to that pseudo-residual. Since the pseudo-residuals are (expected to be) standard normal, a large magnitude value positive or negative is unlikely. To see how the returns affect the pseudo-residuals the magnitude of the returns are plotted over time and they are colored by the magnitude of their respective pseudo-residuals. If the pseudo-residuals are properly identifying outliers, we would expect the magnitude of the pseudo-residual to increase with the magnitude of the return. For poor outlier identification, the magnitude of the return would not have any effect on the pseudo-residual magnitude. A demonstration of the two extremes can be seen in Figure 7. The actual returns data are plotted in Figures 8 and 9. We want to examine how different magnitudes of returns are identified by the two different kinds of pseudo-residuals.



For element pseudo-residuals, the low magnitude returns have low magnitude pseudo-residuals and likewise high magnitude returns have high magnitude pseudoresiduals. So high magnitude returns are being marked as outliers by the element pseudoresiduals. This is particularly apparent in Figure 9. One can see the element pseudo-residual magnitudes increasing with the increase of return magnitudes from the clear color gradient.

For vector pseudo-residuals, there is not as clear of a relationship between return and pseudo-residual. Although smaller returns seem to have smaller vector-pseudo-residuals it is not as consistent as the element pseudo-residuals. A similar trend is observed for the large magnitude returns. The vector pseudo-residuals seem to on average be larger for larger returns, there are many cases where large magnitude returns have small vector-pseudoresiduals. This is can be seen most notably for the highest return magnitude at around day 750. This is the largest drop in stock price that Meta has had in market history. For element pseudo-residuals, it has the largest magnitude pseudo-residual of any other return. The vector pseudo-residual for that day doesn't flag it as that out of the ordinary. There is a similar occurrence for the Intel return around day 375 where a high magnitude return has a high magnitude element pseudo-residual, but a low magnitude vector pseudo-residual. Many more of these cases can be seen throughout the dataset.

So far this analysis has just been visual. We will now use linear regression to see the relationship between the magnitude of pseudoresidual and the magnitude of return.

Let us further investigate the difference between vector and element pseudo-residuals by directly finding their relationship with returns. The behaviour from the returns over time Figures appear to indicate that the farther away the return is from zero, the larger the magnitude of the pseudo-residual. Seen in Figure 10, this is the case for both vector and element pseudo-residuals.

Clearly the negative relationship between the magnitudes of pseudo-residuals and returns is stronger for element pseudo-residuals. The slope is steeper (more positive) for the element pseudo-residuals for all 4 stocks. In addition, the linear fit is better for the element pseudo-residuals. The r squared value for all stocks is significantly higher for the element pseudoresiduals (see Table 3).

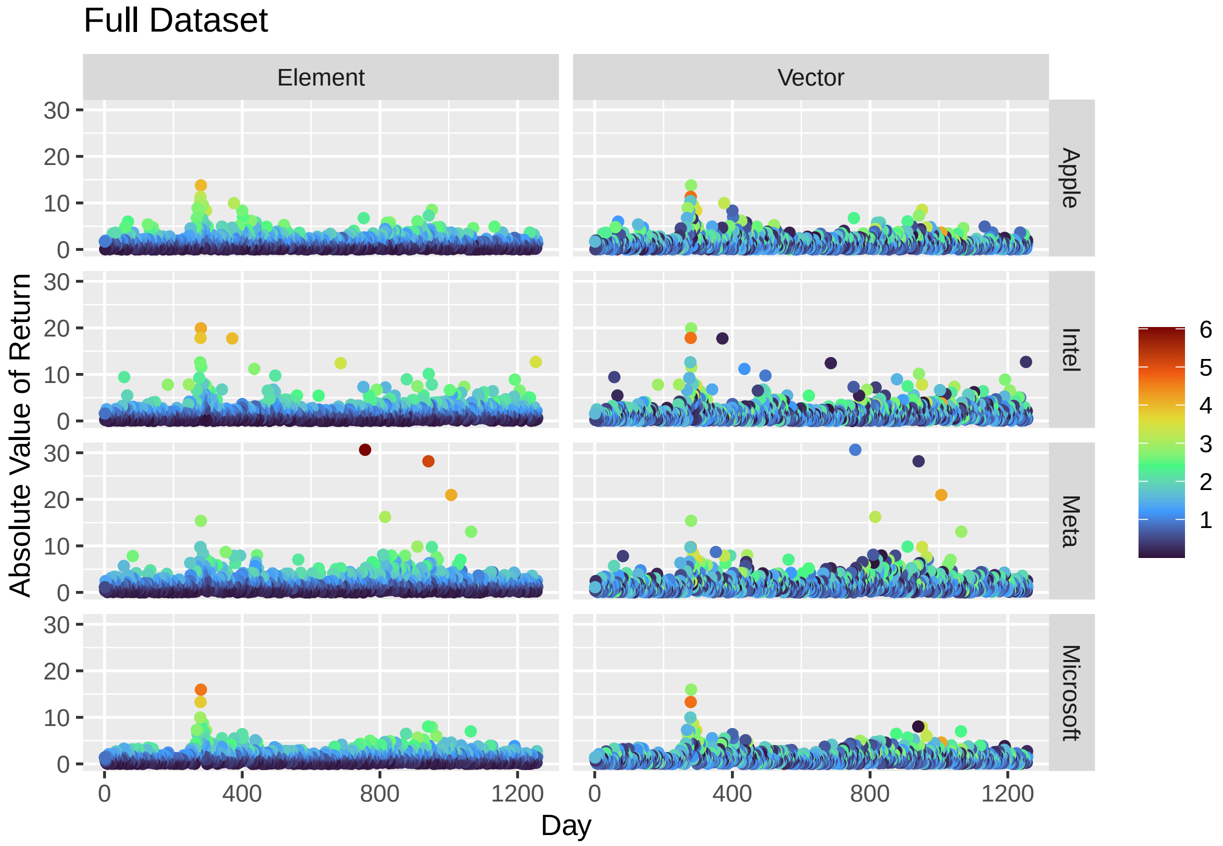


Figure 7: Magnitude of returns for the full dataset plotted against time. The color scale is the magnitude of the element and vector normal pseudo-residuals. From our definition of what outliers are for the returns data, the element pseudo-residuals are significantly better than the vector pseudo-residuals at outlier detection. Draw particular attention to the large magnitude returns with high element pseudo-residuals, but small vector pseudo-residuals. Some of these instances are around day 790 for Meta and around day 375 for Intel. The Meta return is the largest stock drop in Meta's history. It has the largest element pseudo-residual, but it does not have a particularly large vector pseudo-residual. The color scale here is turbo from the `viridis` package, which is a library of colorblind friendly color palletes.

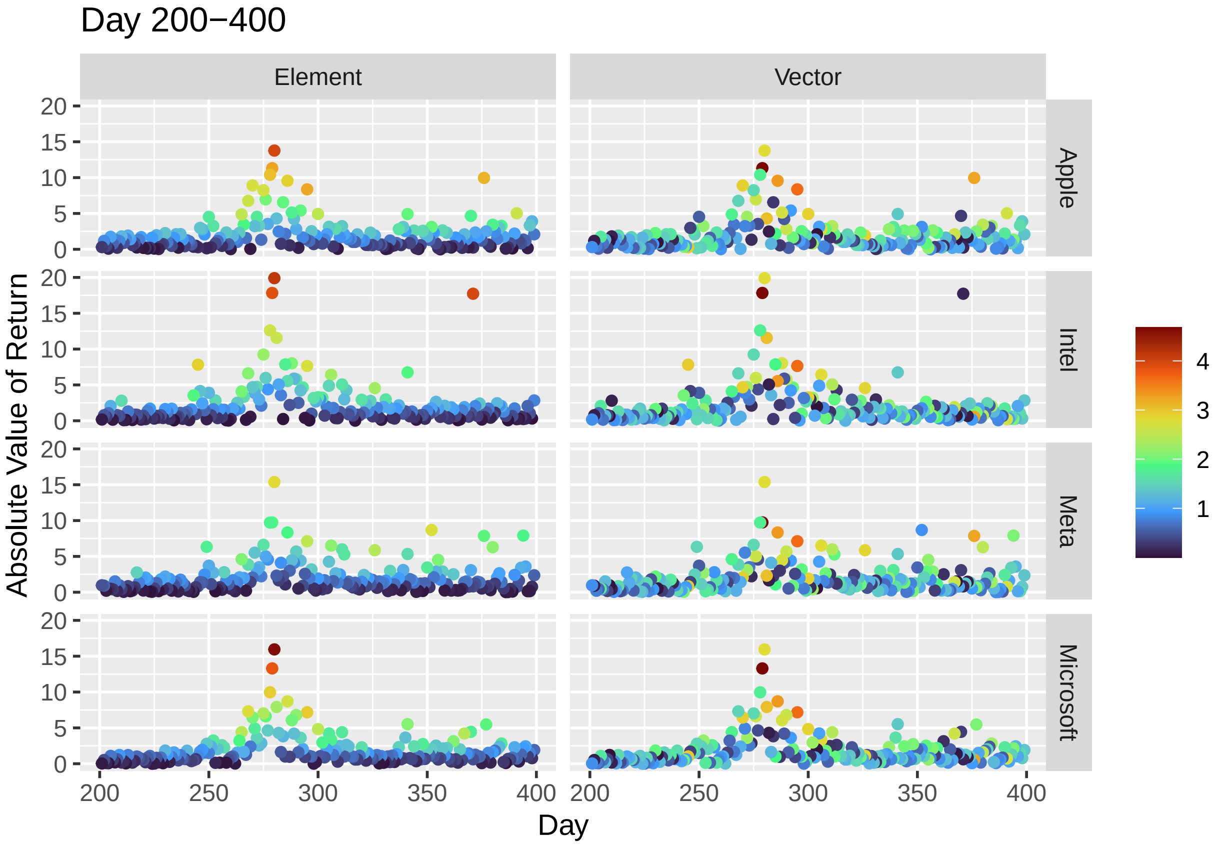


Figure 8: Magnitude of returns for Nov 18 2019 - Sept 3 2020. The magnitude of the element pseudo-residuals seems to increase with the magnitude of the returns. One can see this clearly in the range of return values between day number 250-300. The clear color gradient that was expected (see Figure 7) is seen here for the element pseudo-residuals. Vector pseudo-residuals do not seem to display as strong of a relationship where there is a mix of magnitudes throughout the returns.

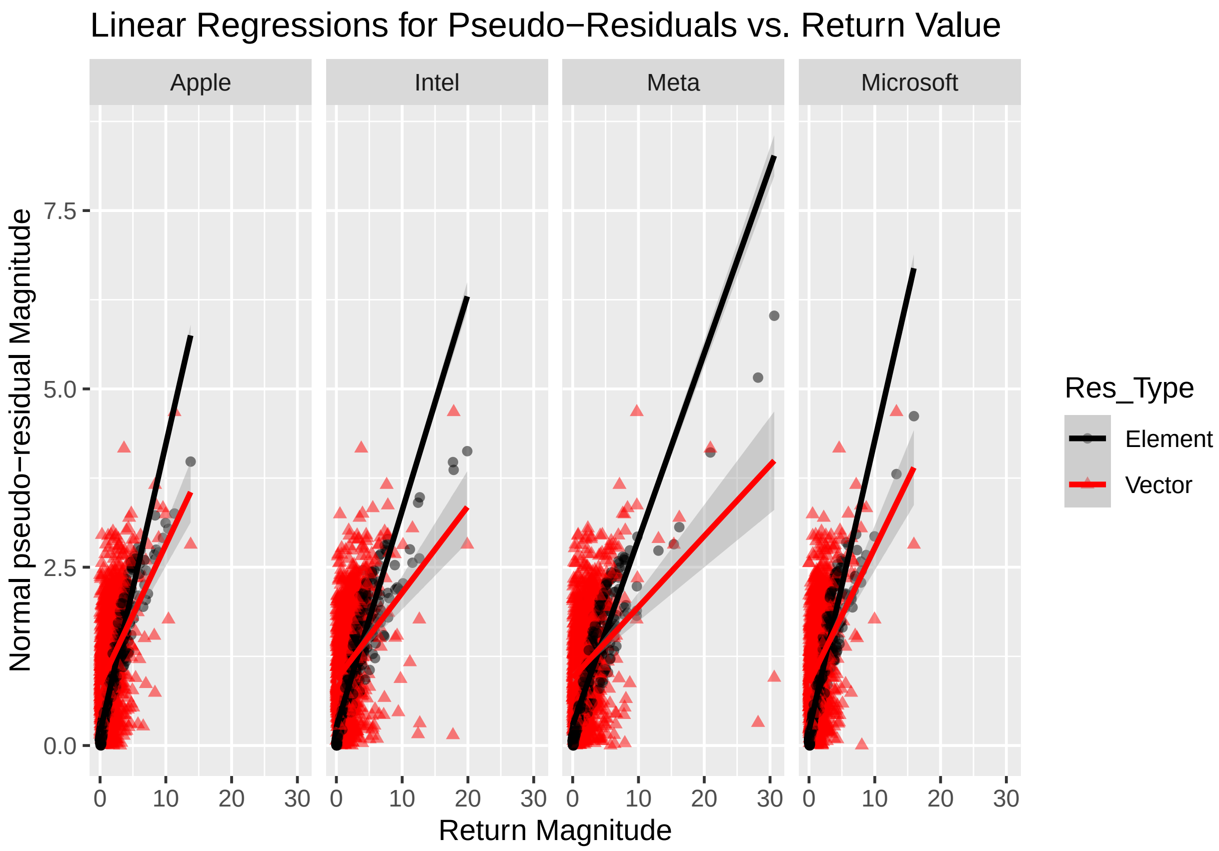


Figure 9: Scatterplot of the normal pseudo-residuals and stock return values seperated by stock with linear models plotted with the normal pseudo-residual as the response variable. The 99% standard error regions are plotted for each line as the gray region for each line. The slope for the element pseudo-residuals is greater for all 4 stocks than the vector pseudo-residuals, meaning that an increase in the magnitude of return is expected to have a greater impact on the element pseudo-residuals than the vector pseudo-residuals. In addition to a greater slope, the element pseudo-residuals seem to fit the linear model a lot better than the vector pseudo-residuals. The element pseudo-residuals are more clustered around the line, wheras the vector pseudoresiduals are more disperse.

Table 2: Element pseudo-residual R Squared Values

Stock	R Squared
Apple	0.9307971
Intel	0.8836330
Meta	0.8642703
Microsoft	0.9119906

Table 3: Vector pseudo-residual R Squared Values

Stock	R Squared
Apple	0.1488362
Intel	0.0950809
Meta	0.0840918
Microsoft	0.1293861

The purpose of this analysis is two fold: Firstly, element pseudo-residuals have been shown to act as expected for outlier detection. In other words, they are behaving similarly to pseudo-residuals for univariate models. This is evidence that the element pseudo-residuals are a multivariate extension of pseudo-residuals for univariate models. Secondly, vector pseudo-residuals are not acting as well in this capacity. The vector pseudo-residual magnitudes do not seem to have as clear of a relationship with the individual stock values. Outlier detection for vector pseudo-residuals is further discussed in appendix C. The outlier detection for the multivariate pseudo-residuals has been examined, let us now examine testing for goodness of fit.

Pseudo-residuals for Testing Goodness of Fit

As previously said, for univariate data, normal pseudo-residuals are expected to have a standard normal distribution for a well fitted model. In order to test this we will be examining the sample means and variances of the pseudoresiduals, as well as doing visual inspection of the density plots against the standard normal. The element and vector pseudo-residual densities can be seen in Figures 11, and 12 respectively.

Table 4: Summary Statistics for Element pseudo-residuals

Stock	Mean	Variance
Apple	-0.0094	0.9515
Intel	-0.0048	0.9688
Meta	-0.0010	0.9557
Microsoft	-0.0042	0.9457

Once again, element pseudo-residuals behave like the pseudo-residuals of their univariate counterparts. Seen in Table 4, the means of the element pseudo-residuals for all four stocks are close to 0, and the variances are close to 1. This is not to say that the model fitted to the returns data set is perfectly fitted. This is to demonstrate that element pseudo-residuals have the behaviour more expected of normal pseudo-residuals.

Table 5: Summary Statistics for Vector pseudo-residuals

Mean	Variance
-0.9674	0.8590891

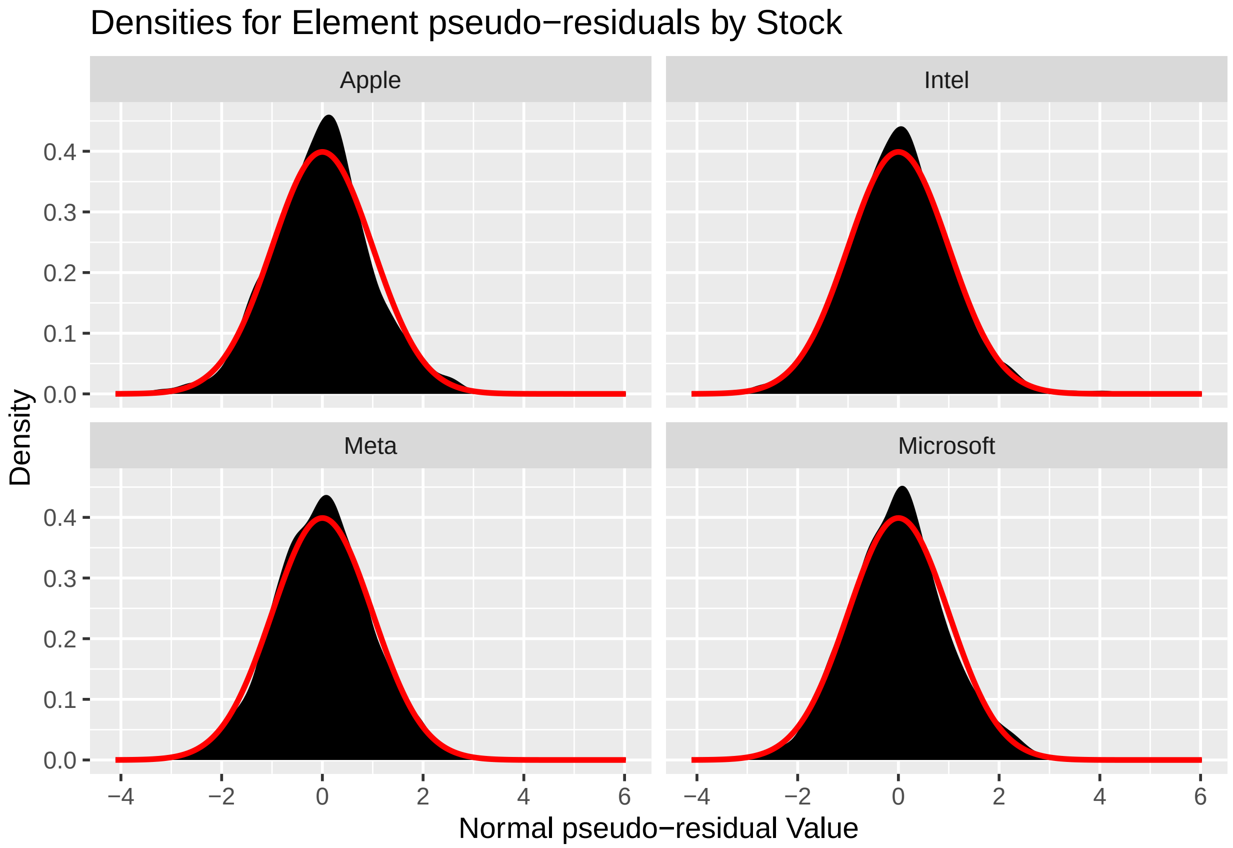


Figure 10: Denisites for the element pseudo-residuals for each stock. Although they do not conform perfectly to the standard normal, they are close. The variance is slightly too small so the peak is higher than that of the standard normal. However, this is not to show that the model fitted for the case study is perfectly well fitted, it is rather to show element pseudo-residuals are exhibiting behavior expected of pseudo-residuals for univariate models.

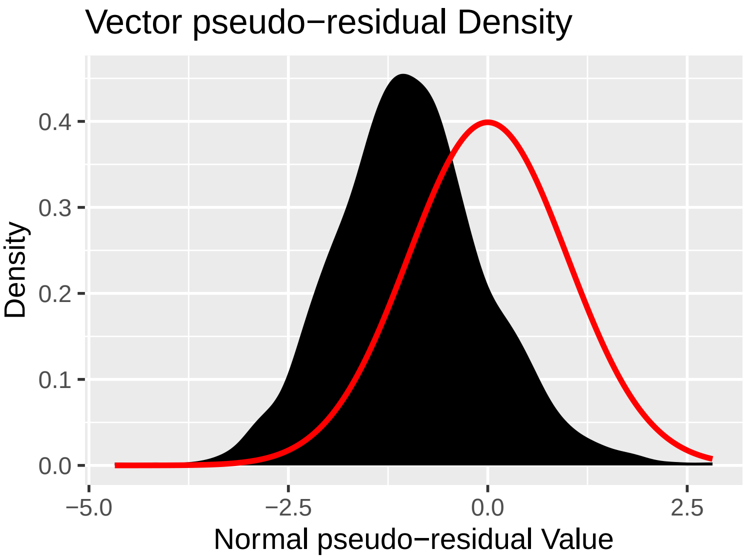


Figure 11: Density plot for the vector pseudo-residuals. Density also appears to be normal, but the mean is shifted left from where one would expect for pseudo-residuals. The normal shape is expected because of the use of the inverse standard normal. The relevent indicator here is whether or not the normal pseudo-residuals are standard normal, not just that they are normally distributed. For the vector pseudo-residuals, there is clear deviation from the standard normal. The mean is significantly far away from 0. Once again vector pseudo-residuals are not behaving like pseudo-residuals for univariate models.

For the vector pseudo-residuals once again there is unexpected behaviour for normal pseudo-residuals. The vector pseudo-residuals are clearly not standard normal distributed. The most clear divergence from standard normality is the sample mean of the vector residuals where its magnitude is much higher than expected for pseudo-residuals for univariate models.

So far, element pseudo-residuals have behaved as expected of normal pseudo-residuals. They identify outliers, and they are close to standard normal. Vector pseudo-residuals have consistently diverged from these behaviours. The question remains, is this a consequence of the returns dataset? To eliminate this possibility, we will now work with simulated data.

Section 5: Simulation Study.

In order to further investigate vector and element pseudo-residuals, simulated data was used in order to eliminate the possibility that the case study dataset was not causing the issue. It is possible that the data set is not suitable to be modelled by HMMs and so model diagnostics used for HMMs on the fitted model have peculiar behaviour. If we instead use data generated from known HMMs we eliminate this possibility, because data that was generated by an HMM should be able to be modelled by an HMM. This is to confirm that element pseudo-residuals are the extension of the previously described normal pseudo-residuals for univariate data, and that vector pseudo-residuals truly do diverge the behaviour of pseudo-residuals for univariate models.

2-Dimensional multivariate-normals were used in order to reduce the number of parameters to be fitted for each trial. The means were set to be close to zero with variances much greater than the means, which was observed in the data in the case study. 200 observations were generated from each model, then a second random model was used as an initial condition for fitting a new model to the generated data. Then the vector and element pseudoresiduals were calculated for the fitted model. Then mean and variance of each kind of pseudo-residual were stored. The model uses 2-dimensional distributions so for each trial, a single mean and variance were stored for the vector pseudo-residuals, and two means and variance were stored for the element pseudo-residual. The results are from the trials where functions were able to run properly. For the simulated data, there were some instances where the model fitting functions did not converge. This only happened with the simulated data, and did not occur very frequently.

Simulated Data Results

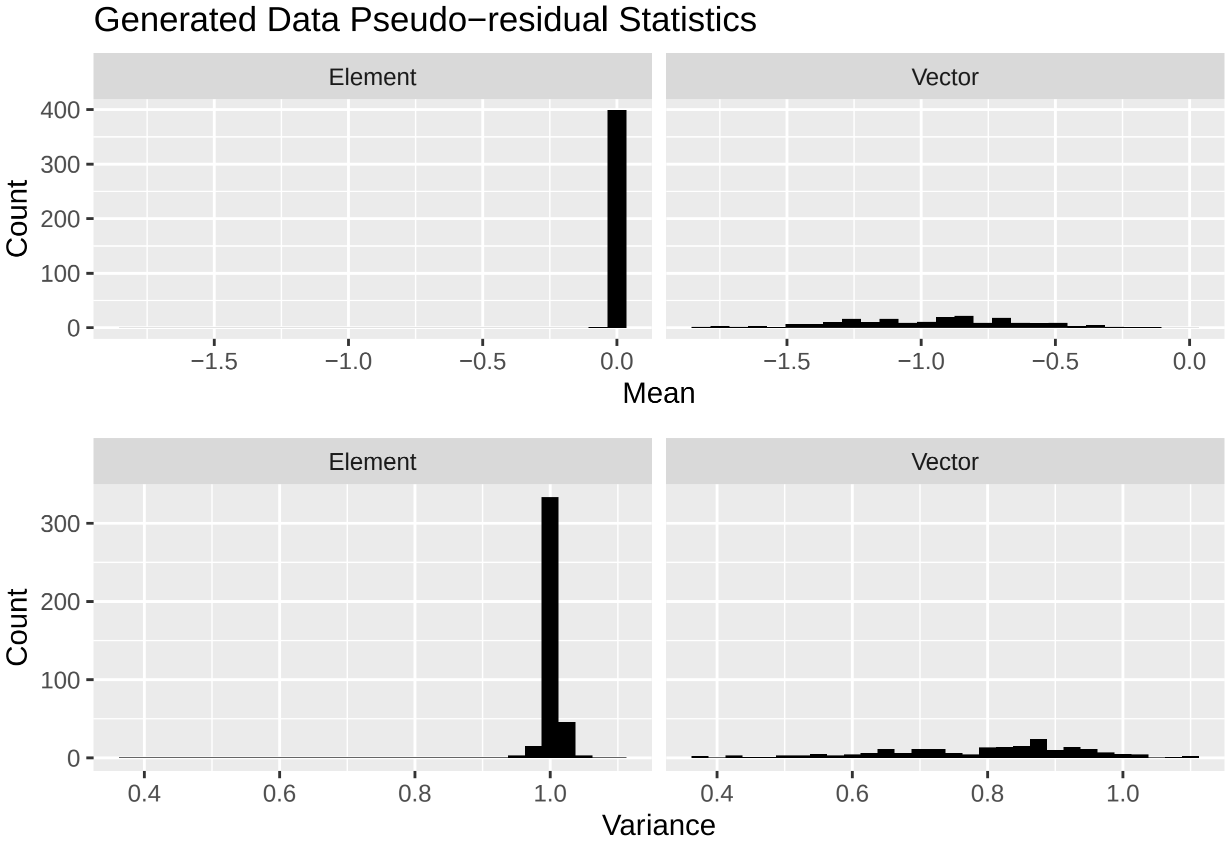


Figure 12: Histograms for statistics of pseudo-residuals of generated data trials. Here the mean and variance of the element and vector pseudo-residuals are used as tests for conformation to the standard normal. The statistics for the element pseudo-residuals indicate that they are predominantly in line with standard normal. There is not a clear conclusion from the vector pseudo-residual statistics aside that they are not standard noraml.

The results of the simulated data, seen in Figure 13, show strong evidence that is was not the data set that was causing the behaviour of the multivariate pseudo-residuals.

The element pseudo-residuals behave the same as seen with the case study. The means and variances of the element pseudo-residuals for the simulated trials were consistently very close to those of the standard normal.

The vector pseudo-residuals once again diverge from univariate pseudo-residual behaviour. The means and variances are widely dispersed and often far from the standard normal values. Unfortunately a clear pattern did not arise from the vector pseudo-residuals. The density of the means and variances of the vector pseudo-residuals for each trial are shown in Figure 14. If the means and variances for the vector pseudoresiduals diverged from standard normal, but perhaps were consistently at a different set of values, then we could start to describe an expected distribution for vector pseudo-residuals. The results however, show no such consistency.

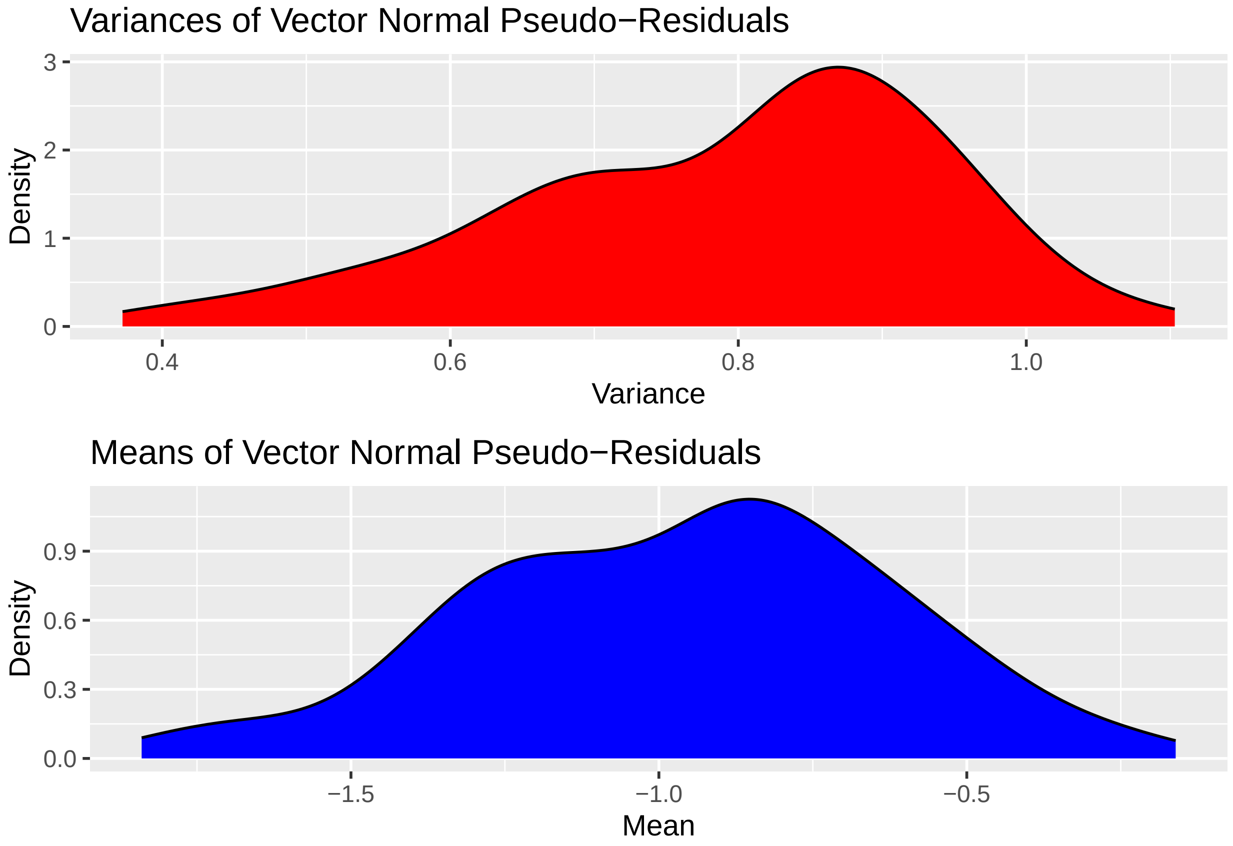


Figure 13: Densities for Means and Variances of vector pseudo-residuals for the generated data trials. Once again there does not seem to be a clear pattern for these values.

Conclusion

In conclusion, two new extensions for normal pseudo-residuals for multivariate data have been described, element and vector pseudo-residuals. Element pseudo-residuals consistently displayed behaviour consistent with pseudo-residuals for univariate data, both identifying outliers, and conforming to the standard normal. Conformation to standard normality was seen not only in the case study, but also for the simulated data trials. Vector pseudo-residuals consistently diverged from both behaviours expected from normal pseudo-residuals, in both the case study and the simulated data trials.

Extensions

There are many avenues to continue exploring the nature of the vector pseudo-residuals. There could be a way to describe the impact of the dimensionality of the state dependent distributions on the behavior of vector pseudo-residuals. The specific distribution (e.g. multivariate normal versus multivariate t) could also have an impact. Describing an expected distribution for vector pseudo-residuals could give them utility in evaluating model fit. For outlier detection, there could be different aspects of the data which the vector pseudo-residuals are better at identifying. They do not seem good for identifying outliers of individual variables of the multivariate data, they could perhaps better describe a statistic of each variable at each time (e.g. the combined mean return of the stocks).

Element pseudo-residuals are an exciting new model diagnostic. Multivariate HMMs are already used in many applications (Martino 2020; Bulla 2012; Southall. 2017). Although this paper has primarily focussed on financial data, the methods are not restricted to financial applications. Having an additional tool model evaluation is good for the many fields using multivariate HMMs. Some more investigation should be done to confirm that element pseudo-residuals maintain their behavior by fitting different kinds of datasets with different numbers of states and different kinds of state-dependent distributions.

Appendix

A: Results from Model Fitting

For completeness sake, here is the resulting model that resulted from the model fitting described above.

Table 6: Means

Apple	Microsoft	Meta	Intel
0.1458	0.1056	0.0548	0.1363
0.0567	0.1014	-0.4038	-0.2420
-1.0593	0.2534	1.0226	-0.4097

Table 7: Variances

Apple	Microsoft	Meta	Intel
1.5149	1.4020	1.6733	1.8198
3.4999	3.4949	4.7567	5.4408
2.3339	0.6887	1.2559	0.4435

Table 8: Correlation State 1

Apple	Microsoft	Meta	Intel
1.0000	0.7677	0.5921	0.6122
0.7677	1.0000	0.5783	0.6758
0.5921	0.5783	1.0000	0.4941
0.6122	0.6758	0.4941	1.0000

Table 9: Correlation State 2

Apple	Microsoft	Meta	Intel
1.0000	0.7640	0.5671	0.5879
0.7640	1.0000	0.5894	0.5916
0.5671	0.5894	1.0000	0.4320
0.5879	0.5916	0.4320	1.0000

Table 10: Correlation State 3

Apple	Microsoft	Meta	Intel
1.0000	0.6752	-0.3192	0.0301
0.6752	1.0000	-0.7581	0.0097
-0.3192	-0.7581	1.0000	-0.4641
0.0301	0.0097	-0.4641	1.0000

Table 11: Transition Probability Matrix

State 1	State 2	State 3
0.90234	0.09031	0.00735
0.44961	0.55039	0.00000
0.52896	0.00000	0.47104

Draw particular attention to correlation matrix of state 3, which has negative correlations.

B: Formula for Conditional Distributions

Presented here is the method for calculating conditional distributions for HMMs (Walter Zucchini 2016). Forward and backward probabilities are important values for calculating many probabilities for HMMs.

Forward Probabilites:

Forward probabilities are defined as follows:

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta} P(\mathbf{x}_1)\boldsymbol{\Gamma} P(\mathbf{x}_2)\boldsymbol{\Gamma} P(\mathbf{x}_3)...\boldsymbol{\Gamma} P(\mathbf{x}_t)$$

Essentially each $\boldsymbol{\alpha}_t$ a vector of the probabilities of observing $\mathbf{x}^{(t)}$ given the state at time t is $1, ..., m$.

$$\boldsymbol{\alpha}_t(j) = \mathbb{P}(\mathbf{X}^{(t)} = \mathbf{x}^{(t)} | c_t = j)$$

Backward Probabilities

Whereas forward probabilities are defined as the probability of observing all the observations up to and including time t, backward probabilities are the probabilities of observing all the observations after time t up to time T. The backward probability vector is defined as follows:

$$\boldsymbol{\beta}_t = \mathbf{\Gamma} P(\mathbf{x}_{t+1}) \mathbf{\Gamma} P(\mathbf{x}_{t+2}) ... \mathbf{\Gamma} P(\mathbf{x}_T) \mathbf{1}$$
$$\boldsymbol{\beta}_T = \mathbf{1} = \{1, 1, ..., 1\} \quad (m \text{ times})$$

These values will be used to define the conditional probabilities for each observation. The pdf for each observation is going to be the probability of an observation conditioned on the rest of the observations. To be explicit:

$$\mathbb{P}[X_t = \mathbf{x} | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}] = \mathbb{P}[\frac{X_t = \mathbf{x}, \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}}{\mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}}]$$

The numerator is just the likelihood L_T as described above, just with the observation at time t replaced with the generic observation vector $\mathbf{x}$. The above probability can be calculated using the following values:

$$\mathbb{P}(\mathbf{X}_t = \mathbf{x} | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}) = \frac{\boldsymbol{\delta} P(\mathbf{x}_1) \mathbf{\Gamma} P(\mathbf{x}_2) ... \mathbf{\Gamma} P(\mathbf{x}_{t-1}) \mathbf{\Gamma} P(\mathbf{x}) ... \mathbf{\Gamma} P(\mathbf{x}_T) \mathbf{1}'}{\boldsymbol{\delta} P(\mathbf{x}_1) \mathbf{\Gamma} P(\mathbf{x}_2) \mathbf{\Gamma} P(\mathbf{x}_3) ... \mathbf{\Gamma} P(\mathbf{x}_{t-1}) \mathbf{\Gamma} P(\mathbf{x}_{t+1}) ... \mathbf{\Gamma} P(\mathbf{x}_T) \mathbf{1}'}$$

This can be simplified using our notation from above for forward and backward probabilities to then get the pdf for an observation:

$$\mathbb{P}(\mathbf{X}_t = \mathbf{x} | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}) = \frac{\boldsymbol{\alpha}_{t-1} \mathbf{\Gamma} P(\mathbf{x}) \boldsymbol{\beta}_t'}{\boldsymbol{\alpha}_{t-1} \mathbf{\Gamma} \boldsymbol{\beta}_t'}$$

From here, getting the conditional cumulative distribution function is trivial, one need only calculate the cdf for $P(\mathbf{x})$ instead of the pdf (Essentially `pmvnorm` instead of `dmvnorm`). Here $\mathbf{P}(\mathbf{x})$ indicates the use of the cdf instead of the pdf.

$$\mathbb{P}(\mathbf{X} \leq \mathbf{x} | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}) = \frac{\boldsymbol{\alpha}_{t-1} \mathbf{\Gamma} \mathbf{P}(\mathbf{x}) \boldsymbol{\beta}_t'}{\boldsymbol{\alpha}_{t-1} \mathbf{\Gamma} \boldsymbol{\beta}_t'}$$

Code

The code written for this project can be found at https://github.com/Tazman-Libson/St_Andrews_Final_Project_2024

References:

Boruah, Saptarshi, and Subhash Basishtha. 2013. "A Study on HMM Based Speech Recognition System." In *2013 IEEE International Conference on Computational Intelligence and Computing Research*, 1–5. https://doi.org/10.1109/ICCIC.2013.6724147.

Bulla, Lagona, J. 2012. *A Multivariate Hidden Markov Model for the Identification of Sea Regimes from Incomplete Skewed and Circular Time Series.* JABES. https://doi.org/10.1007/s13253-012-0110-1.

Chang, Winston. 2023. *Cookbook for r.* http://www.cookbook-r.com/Graphs/.

Conners, Michelot, M. G. 2021. *Hidden Markov Models Identify Major Movement Modes in Accelerometer and Magnetometer Data from Four Albatross Species.* Movement Ecology. https://doi.org/10.1186/s40462-021-00243-z.

Fernández-Fontelo, Cabaña, A. 2019. *Untangling Serially Dependent Underreported Count Data for Gender-Based Violence.* Statistics in Medicine. https://doi.org/10.1002/sim.8306.

Francois, Romain, and Diego Hernangómez. 2023. *Bibtex: Bibtex Parser.* https://CRAN.R-project.org/package=bibtex.

Friendly, Michael, John Fox, and Phil Chalmers. 2022. *Matlib: Matrix Functions for Teaching and Learning Linear Algebra and Multivariate Statistics.* https://CRAN.R-project.org/package=matlib.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, et al. 2023. *viridis(Lite) - Colorblind-Friendly Color Maps for r*. https://doi.org/10.5281/zenodo.4679423.

Genz, Alan, and Frank Bretz. 2009. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Heidelberg: Springer-Verlag.

Kwak, & Kim, S. K. 2017. *Statistical Data Preparation: Management of Missing Values and Outliers*. Korean journal of anesthesiology. https://doi.org/10.4097/kjae.2017.70.4.407.

Lintern, Kho, A. 2023. *Shifts in Stream Salt Loads During and After Prolonged Droughts. Hydrological Processes*. https://doi.org/10.1002/hyp.14901.

Lumley, Thomas. 2022. *Dichromat: Color Schemes for Dichromats*. https://CRAN.R-project.org/package=dichromat.

Martino, Guatteri, A. 2020. *Multivariate Hidden Markov Models for Disease Progression. Statistical Analysis and Data Mining*. The ASA Data Science Journal. https://doi.org/10.1002/sam.11479.

McClintock, & Lander, B. T. 2024. *A Multistate Langevin Diffusion for Inferring Behavior-Specific Habitat Selection and Utilization Distributions*. Ecological Society of America. https://doi.org/10.1002/ecy.4186.

NASDAQ. 2024. https://www.nasdaq.com/market-activity/stocks/.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

Oelschläger, Lennart, and Timo Adam. 2023. "Detecting Bearish and Bullish Markets in Financial Time Series Using Hierarchical Hidden Markov Models." *Statistical Modelling* 23 (2): 107–26. https://doi.org/10.1177/1471082X211034048.

Oelschläger, Lennart, Timo Adam, and Rouven Michels. 2024. *fHMM: Fitting Hidden Markov Models to Financial Data*. https://CRAN.R-project.org/package=fHMM.

Osborne, A., J. W. & Overbay. 2004. *The Power of Outliers (and Why Researchers Should ALWAYS Check for Them)*. Practical Assessment, Research,; Evaluation. https://doi.org/10.7275/qf69-7k43.

Privault, Nicolas. 2013. *Understanding Markov Chains*. https://doi-org.ezproxy.st-andrews.ac.uk/10.1007/978-981-4451-51-2.

R Core Team. 2023b. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.RR-project.org.

———. 2023a. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Silvey, S. D. 1975. *Statistical Inference (1st Ed.)*. Routledge. https://doi-org.ezproxy.st-andrews.ac.uk/10.1201/9780203738641.

Southall., Stacy L. DeRuiter. Roland Langrock. Tomas Skirbutas. Jeremy A. Goldbogen. John Calambokidis. Ari S. Friedlaender. Brandon L. 2017. *A Multivariate Mixed Hidden Markov Model for Blue Whale Behaviour and Responses to Sound Exposure*. https://doi.org/10.1214/16-AOAS1008.

Tobias Preis, H. Eugene Stanley, Dror Y. Kenett. 2012. *Quantifying the Behavior of Stock Correlations Under Market Stress*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3475344/#b1.

Walter Zucchini, Roland Langrock, Iain L. MacDonald. 2016. *Hidden Markov Models for Time Series*. https://doi-org.ezproxy.st-andrews.ac.uk/10.1201/b20790.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2019. "TinyTeX: A Lightweight, Cross-Platform, and Easy-to-Maintain LaTeX Distribution Based on TeX Live." *TUGboat* 40 (1): 30–32. https://tug.org/TUGboat/Contents/contents40-1.html.

———. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

———. 2024. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*.

https://github.com/rstudio/tinytex.