Synthetic Likelihoods for Parameter Inference in Electricity Spot Price Models

Jake Ireland

Supervised by Dr. Matteo Fasiolo
Level H/6
20 Credit Points

May 7, 2022

Acknowledgement of Sources

For all ideas taken from other sources (books, articles, internet), the source of the ideas is mentioned in the main text and fully referenced at the end of the report.

All material which is quoted essentially word-for-word from other sources is given in quotation marks and referenced.

Pictures and diagrams copied from the internet or other sources are labelled with reference to the web page or book, article, etc.

Signed: Jake Ireland

Date: May 7, 2022

To my parents, Paul and Briony, for always supporting me and lighting my spark in mathematics.
To my sister, Beth, for inspiring me with your incredible work ethic and dedication.
To my partner, Zach, for your love, your encouragement and your companionship in our quest for a just and progressive world.

Contents

List of Figures

List of Tables

1 Introduction

This project uses 'Synthetic Likelihoods' (Wood 2010) to perform parameter inference for electricity spot price models. The regime-switching model proposed in Huisman & Mahieu (2003) will be used to model the spot price. Additionally, this project will evaluate the use of the Robust Covariance Matrix (Huber 1967) as an estimate of the sample covariance calculated during the evaluation of a Synthetic Likelihood. Computational benefits will be explored and the resulting fits compared.

Synthetic Likelihoods were first introduced as a method of statistical inference for chaotic ecologial dynamical systems (Wood 2010). In these systems, slight changes to model parameters have large effects on the resulting trajectory (May 1976). This makes the probability density function unsuitable for determining statistical fit. Thus, traditional likelihood approaches such as Maximum Likelihood Estimation do not work (Wood 2010). However, in Wood (2010), this is solved by a) choosing a vector of summary statistics that characterise the underlying dynamics of the model and b) maximising the likelihood of the observed data presenting these statistics.

Spot electricity prices present frequent extreme observations or price spikes (Pilipović 2007, p. 21). These spikes are subject to the 'double penalty' effect (Haben et al. 2014). This means that sample trajectories produced even from the same parameters can look completely different. Thus, traditional likelihood-based inference is impossible. This project will show how Synthetic Likelihoods can elegantly overcome this obstacle.

The Robust Covariance Matrix was proposed as an estimate of the asymptotic covariance of the Maximum Likelihood Estimator when the underlying model is misspecified (Huber 1967). However, it will be shown that the Robust Covariance Matrix can also be used within the Synthetic Likelihood framework to provide significant computational benefit.

The structure of this project is as follows. In Section 2, a model for spot electricity prices will be explored. Then, in Section 3, Synthetic Likelihoods will be introduced alongside the Robust Covariance Matrix. In Section 4, the electricity spot price model will be fit using Synthetic Likelihoods. Finally, in Section 5, the goals and results of the project will be discussed.

2 Electricity Spot Price Model

Spot prices demonstrate several features which will need to be included in the model. Specifically, 'high volatility, mean-reversion, seasonality, and frequent extreme jumps' (Huisman & Mahieu 2003). Mean reversion comes from the existence of a long-term equilibrium which fluctuations return to (Pilipović 2007, p. 26). Seasonality is caused by a fundamental change in the equilibrium price caused by heating/cooling demand in the Winter/Summer months (Pilipović 2007, p. 30-31). High volatility and price spikes are caused by fluctuations in demand combined with low elasticity of supply since electricity cannot be readily stored (Pilipović 2007, p. 26). To account for these features, Huisman & Mahieu (2003) proposes the following regime-switching model.

2.1 Model Specification

Let $s(t)$ denote the natural log of the spot price on day t where $t \in \mathbb{Z}_{>0}$. Then, $s(t)$ is modelled as the sum of a deterministic component $f(t)$ and stochastic component $x(t)$. The deterministic component is given as

$$f(t) = \mu_0 + \beta_1 D_1(t) + \beta_2 D_2(t) \tag{1}$$

where μ_1, β_1 and β_2 are parameters in $\mathbb{R}$ and $D_1(t)$ and $D_2(t)$ are dummy variables representing whether day t is a Saturday or Sunday respectively. This accounts for an intraweek demand profile with lower prices during weekends.

The stochastic component is where the regimes come in. The model in Huisman & Mahieu (2003) includes three regimes. A regime for normal price levels (regime 0), a regime for a price spike (regime 1) and a regime for reverting after a price spike (regime -1). Regimes 0 and -1 are specified the same. The inclusion of regime -1 allows for proper specification of the mean reversion component. That is, there may be a stronger reverting process after a spike compared to during regime 0 (Huisman & Mahieu 2003). To move between regimes, a Markov process on the state space $\{0, +1, -1\}$ with transition diagram given in Figure 1, is used.

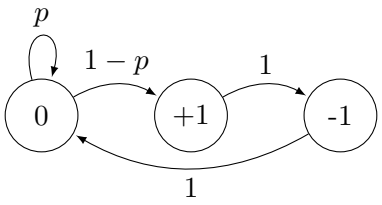

Figure 1: State Transition Diagram

The model starts in regime 0 (default price levels), then with some probability $1 - p$ it moves to regime 1 (the price spike regime). The day after entering regime 1 it enters regime -1 (the reverting regime). Then, the day after that, it returns to regime 0.

The regimes are specified in terms of the day-on-day change in the stochastic component $\mathrm{d}x(t) := x(t) - x(t-1)$ with $x(1) = 0$. Let (ϵ_t) be a sequence of independent standard Normal random variables. For normal price levels (regime 0) model

$$\mathrm{d}x(t) = -\alpha_0 x(t-1) + \sigma_0 \epsilon_t \tag{2}$$

where $\alpha_0 \in (0, 1)$ and $\sigma_0 \in (0, \infty)$ are parameters. Parameter α_0 specifies the level of mean reversion. That is, the greater α_0 is, the quicker any fluctuations away from the deterministic component will be restored.

The reverting regime (regime -1) is modelled exactly the same. Model

$$\mathrm{d}x(t) = -\alpha_{-1} x(t-1) + \sigma_{-1} \epsilon_t \tag{3}$$

where $\alpha_{-1} \in (0, 1)$ and $\sigma_{-1} \in (0, \infty)$ are parameters. Finally for the spike regime (regime 1) a Normally distributed price spike is introduced as

$$\mathrm{d}x(t) = \mu_1 + \sigma_1 \epsilon_t \tag{4}$$

where $\mu_1 \in \mathbb{R}$ and $\sigma_1 \in (0, \infty)$ are parameters.

This completes the specification of the model. The next section will look at how the constrained parameters in the model can be handled.

2.2 Constrained Parameters

In the model, there are several constrained parameters. These will be handled by introducing unconstrained transformed parameters. Consider the functions $f(x) = e^x$ and $g(x) = \frac{e^x}{1-e^x}$. Both functions are defined on $\mathbb{R}$ and are invertible. Further, the image of f is $(0, \infty)$ and the image of g is $(0, 1)$. Thus, for the parameters constrained to $(0, \infty)$, unconstrained parameters τ_{σ_i} where $\sigma_i = f(\tau_{\sigma_i})$ for $i = 0, 1, -1$ can be used. Similarly, unconstrained parameters τ_p and τ_{α_j} where $p = g(\tau_p)$ and $\alpha_j = g(\tau_{\alpha_j})$ for $j = 0, -1$ can be used for the parameters constrained to $(0, 1)$. Now, the task is to infer unconstrained parameters: $\mu_0, \beta_1, \beta_2, \tau_{\alpha_0}, \tau_{\sigma_0}, \mu_1, \tau_{\sigma_1}, \tau_{\alpha_{-1}}, \tau_{\sigma_{-1}}$ and τ_p.

Before introducing Synthetic Likelihoods, it is beneficial to look at some example trajectories to better understand of how varying certain parameters affect the model.

2.3 Example Trajectories

As will be seen in Sections 3 and 4, the process of choosing statistics for Synthetic Likelihoods is a creative and visual process that requires an understanding of how the model parameters affect the resulting trajectory. This section will provide intuition for the mean reversion parameter, α_0, as well as the spike parameters, μ_1 and p. The fitted parameters from Huisman & Mahieu (2003) for the UK Power Market (Table 1) will be used as defaults; values differing from Table 1 will be noted.

Table 1: Fitted Parameter Values for UK Telerate Market (Huisman & Mahieu 2003)

Parameter:	μ_0	β_1	β_2	α_0	σ_0	μ_1	σ_1	α_{-1}	σ_{-1}	p
Value:	2.852	-0.089	-0.192	0.112	0.144	0.103	0.542	0.313	0.453	0.95

First, consider the mean reversion parameter α_0. In Figure 2 it can be seen that setting $\alpha_0 = 0$ makes the resulting trajectory smoother and there appears to be a downwards trend that would have been rectified by introducing some mean-reversion. When it comes to using Synthetic Likelihoods, as will be seen in Section 3, statistics need to be chosen that represent the dynamics of the model. For example, the plots in Figure 2 suggest different relationships between $s(t)$ and $s(t-1)$. Thus, autocorrelation coefficients may be useful statistics to include.

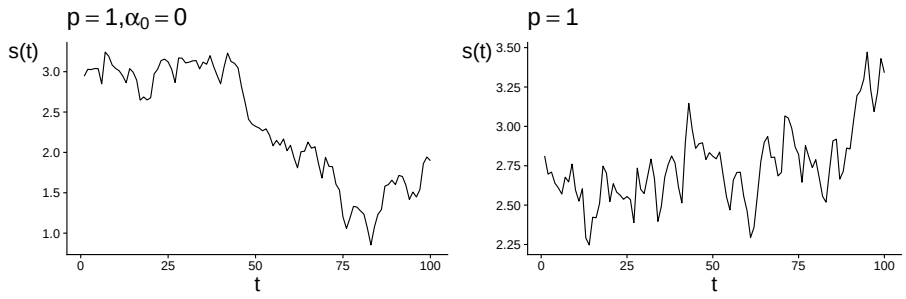

Figure 2: Understanding the Mean Reversion Parameter

Moving to the spike parameters. Figure 3 shows graphs with different parameters for μ_1 and p. Clearly, μ_1 controls the average height of the spikes and p the frequency of spikes. Statistics that capture the effects of these parameters will be presented in Section 4.2.1.

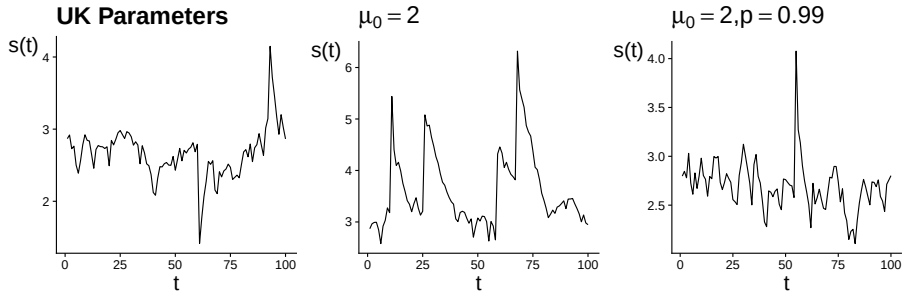

Figure 3: Understanding the Spike Parameters

Now that a model that can capture the nuance of electricity prices has been specified, the next section will introduce Synthetic Likelihoods. From this point forward, the model put forward in this section will be referred to as Model $\mathcal{A}$.

3 Synthetic Likelihoods

There are two key motivators for exploring the use of Synthetic Likelihoods in this set-up. The first is that Model $\mathcal{A}$ has no (unconditioned) tractable likelihood function. This is caused by the regime-switching nature of the model. Following Huisman & Mahieu (2003), separate likelihood functions could be obtained by conditioning on each regime.

Then, the Kalman Filter could be applied to obtain Maximum Likelihood Estimates. However, there are good reasons to seek a method which does not require the Kalman Filter.

One reason is that the Kalman Filter puts restrictions on how the individual regimes can be specified. In particular, the Kalman Filter requires the model to be written in the form

$$\begin{cases} \boldsymbol{x}(t+1) = \boldsymbol{\Phi}(t+1;t)\boldsymbol{x}(t) + \boldsymbol{u}(t) \\ \boldsymbol{y}(t) = \boldsymbol{M}(t)\boldsymbol{x}(t) \end{cases} \tag{5}$$

where $\boldsymbol{x}(t)$ is the state of the system, $\boldsymbol{\Phi}(t+1;t)$ is a transition matrix, $\boldsymbol{u}(t)$ is a zero-mean Normal random variable, $\boldsymbol{M}(t)$ is a matrix and $\boldsymbol{y}(t)$ is the output of the system; all at time t (Kalman 1960). For Model $\mathcal{A}$, it means the distribution of the spikes must be Normal (Equation 4). This is an issue since the spikes should always be positive which under a Normal distribution will not always be the case. It is more appropriate to replace the Normal distribution with the Log-Normal distribution. This is possible with Synthetic Likelihoods but not with the Kalman Filter. In short, Synthetic Likelihoods offer the opportunity to expand upon Model $\mathcal{A}$ in useful and meaningful ways.

The second motivator for using Synthetic Likelihoods is caused by the double penalty effect or 'double penalty problem'. The classical 'double penalty problem' refers to the modelling of rainfall. Specifically, where a model that accurately predicts intensity, size and timing but is off on location incurs a 'double penalty' with respect to point-wise error measures (Keil & Craig 2009).

In Model $\mathcal{A}$, the double penalty effect occurs in the following sense. First, the regime that the model is in on day t is a hidden state. Thus, it is not known when the price spikes occur. Consider Figure 4. There are two trajectories. The observed trajectory and a simulated trajectory. The simulated trajectory has spikes of a similar magnitude and frequency to the spikes of the observed trajectory. Perhaps the parameters used to produce the simulation are good parameters? If one were to successively sample with these parameters, one would imagine the situation displayed in Figure 4 would be common. The simulation would have approximately the right number and size of spikes, but they would rarely coincide. This is the double penalty effect.

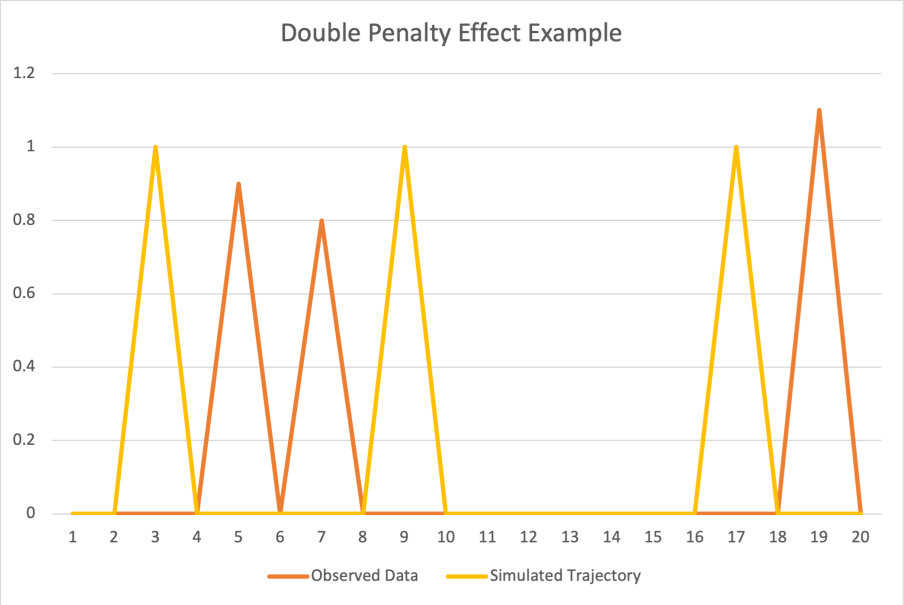

Figure 4: Double Penalty Effect (Haben et al. 2014)

One solution would be to integrate out the hidden state from the joint density of the observations and the hidden state, as with the Kalman Filter. However, this is undesirable as it restricts how Model $\mathcal{A}$ can be specified. There may be particle filter algorithms that could accomplish this without restricting the specification of model $\mathcal{A}$, but they are computationally intensive and beyond the scope of this project. Thus, a method of parameter inference that does not require integrating out the hidden state is required.

Synthetic Likelihoods, as will be seen in Section 3.1, do not need a tractable likelihood function. Further the 'double penalty' effect can be solved by only choosing statistics that are agnostic to the locations of the price spikes. There is no need to integrate out the hidden state. Now there is sufficient motivation, Synthetic Likelihoods can be introduced.

3.1 Synthetic Likelihood Estimation

Synthetic Likelihood Estimation will be introduced in comparison to Maximum Likelihood Estimation. With Maximum Likelihood Estimation parameters are chosen to maximise the chance of producing the observed trajectory (Rossi 2018, p. 226). With Synthetic Likelihood Estimation, one first selects summary statistics that characterise

the dynamics of the model, then chooses parameters that maximise the chance of producing the observed summary statistics (according to the log-likelihood of a multivariate Normal distribution).

Crucially, there is no need for a tractable likelihood function and the double penalty effect need not apply. The only requirements are to be able to sample from the model and choose appropriate summary statistics.

The method is precisely defined in Algorithm 1 (Fasiolo et al. 2016).

Algorithm 1 Calculation of the Synthetic Likelihood

Let $p(\boldsymbol{y}|\boldsymbol{\theta})$ be a statistical model
Let $\boldsymbol{y}^0$ be the observed trajectory
Let $S(\cdot)$ be a function that computes the summary statistics of a trajectory
Let $\boldsymbol{s}^0 = S(\boldsymbol{y}^0)$
Assume $S(\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$
Let $\varphi(\cdot)$ be the log-likelihood function of a multivariate Normal distribution

procedure SYNTHETIC_LIKELIHOOD($p, S, N, \boldsymbol{s}^0$)
    Simulate N data sets $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^N$ from the model $p(\boldsymbol{y}|\boldsymbol{\theta})$
    Calculate $\boldsymbol{s}^i = S(\boldsymbol{y}^i)$ for $1 \leq i \leq N$
    Calculate the sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and sample covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ of $\boldsymbol{s}^1, \ldots, \boldsymbol{s}^N$
    Estimate the Synthetic Likelihood $\hat{p}(\boldsymbol{s}^0|\boldsymbol{\theta}) = \varphi(\boldsymbol{s}^0|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$
    return $\hat{p}(\boldsymbol{s}^0|\boldsymbol{\theta})$
end procedure

To explore Synthetic Likelihood Estimation in more detail it will be used to infer the parameters of a Zero-Inflated Gamma distribution.

3.2 Example: Zero-Inflated Gamma

First, the Zero-Inflated Gamma random variable needs defining.

Definition 3.1. Let $G \sim \text{Gamma}(\alpha, \beta)$, a Gamma random variable with shape parameter α and rate parameter β. Let $B \sim \text{Bernoulli}(p)$ where

$$p = \mathbb{P}(B = 1) = 1 - \mathbb{P}(B = 0).$$

Then, X = (1-B)G, is a Zero-Inflated Gamma random variable with shape parameter α,

rate parameter β and zero parameter p. In this case, write $X \sim \text{ZIG}(\alpha, \beta, p)$.

With this definition, one can perform Synthetic Likelihood Estimation.

Let $X \sim \text{ZIG}(\alpha, \beta, p)$. For this example, choose the true parameters to be $\alpha^* = 3$, $\beta^* = 1$ and $p^* = 0.3$. Next, generate a sample of $m = 2000$ observations. This will be the observed trajectory denoted $\boldsymbol{y}^0$. To construct the Synthetic Likelihood, summary statistics need to be chosen. It makes sense to use the sample mean and standard deviation since, for a Gamma distribution, these can be calculated in terms of α and β. It also makes sense to include the proportion of zeros to correctly inference the true value of p. Let $\bar{x}(\boldsymbol{y}^i)$ denote the sample mean and $\text{sd}(\boldsymbol{y}^i)$ the standard deviation of trajectory $\boldsymbol{y}^i$. So for trajectory $\boldsymbol{y}^i$, compute

$$S(\boldsymbol{y}^i) = (\bar{x}(\boldsymbol{y}^i), \text{sd}(\boldsymbol{y}^i), \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\{y_j^i = 0\}). \tag{6}$$

To estimate the Synthetic Likelihood for a given $\boldsymbol{\theta} = (\alpha, \beta, p)$, do the following. Take $N = 100$ samples and transform each sample to a summary statistic by $S(\cdot)$. Then, calculate the sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and the sample covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ of the summary statistics. Finally, compute $\hat{p}(\boldsymbol{s}^0|\boldsymbol{\theta}) = \varphi(\boldsymbol{s}^0|\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$ where $\boldsymbol{s}^0 = S(\boldsymbol{y}^0)$.

An estimate of the Synthetic Likelihood can now be computed for any $\boldsymbol{\theta}$. To compute Synthetic Likelihood Estimate $\hat{\boldsymbol{\theta}}$, a stochastic optimiser is needed to maximise the Synthetic Likelihood over all $\boldsymbol{\theta}$. The 'ML' optimiser from *Synlik* (Matteo Fasiolo and Simon N. Wood 2014) will do this. Implementing the above in R will result in convergence to the true parameters. This is shown in Figure 5.

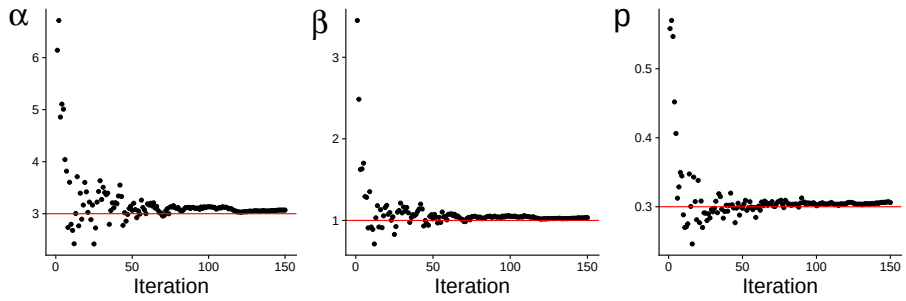

Figure 5: Convergence to True Parameters (ZIG no RCM)

Synthetic Likelihood Estimation is a computationally intensive operation which is due

to the large number of samples that need to be generated to evaluate the likelihood at each stage of the optimisation. Indeed, in this example, 50 Synthetic Likelihoods were simulated at each stage of the optimisation. Further, each Synthetic Likelihood required 100 samples to generate a likelihood. Thus, the entire optimisation generated $150 \times 50 \times 100 = 750000$ samples. This does not pose a problem for the ZIG example as the number of statistics is small. However, as the number of statistics grows, this becomes a significant issue as more samples are required to obtain a less noisy estimate of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. In the next section, the Robust Covariance Matrix (RCM) will be introduced. This will reduce the number of samples required per likelihood to *one*.

3.3 Robust Covariance Matrix

To motivate this section notice that, in Algorithm 1, $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ could be estimated by simply evaluating the statistics over a sufficiently long trajectory. The Robust Covariance Matrix can be used in the same way to estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. Thus, only one sample is needed to produce a likelihood. However, in order to use the Robust Covariance Matrix, only a certain class of statistics can be used - M-estimators (Huber 1967).

Definition 3.2. An M-Estimator is a solution θ that minimises $L = \sum_{i=1}^{n} \rho(x_i, \theta)$ where ρ is an arbitrary function.

In Section 3.8, it will be shown that M-estimators are asymptotically Normal. As an example of an M-Estimator, it will be shown that any Maximum Likelihood Estimator is an M-Estimator.

Proposition 3.3. *Any Maximum Likelihood Estimator is an M-Estimator.*

Proof. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with density $f(x|\theta)$ where θ is a parameter. Let $x_1, x_2, \ldots, x_n$ be the realisations of those respective random variables. Then, the Maximum Likelihood Estimator for θ is given by

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left[\log \prod_{i=1}^{n} f(x_i|\theta) \right] = \underset{\theta \in \Theta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} -\log f(x_i|\theta) \right].$$

Thus, taking $\rho(x_i, \theta) = -\log f(x_i|\theta)$, one obtains that the Maximum Likelihood Estimator of θ is an M-Estimator. □

In order to define the Robust Covariance Matrix, the sample covariance is needed. The

definition below is from Krzanowski (2000).

Definition 3.4. Let $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ be a matrix where $(\boldsymbol{M})_{ij}$ is the i'th observation of the j'th variable. Obtain matrix $\boldsymbol{Q}$ by subtracting the sample column means of $\boldsymbol{M}$ from each column of $\boldsymbol{M}$. Then, obtain the sample covariance as

$$\text{cov}(\boldsymbol{M}) = \frac{1}{n-1}\boldsymbol{Q}^T\boldsymbol{Q}.$$

In Definition 3.4, $\boldsymbol{Q}^T$ represents the transpose of matrix $\boldsymbol{Q}$.

Now the Robust Covariance Matrix can be defined. This project will modify the definition of the Robust Covariance Matrix given in Huber (1967).

Definition 3.5. Let $\boldsymbol{y}^1$ be a sample trajectory with length m. Let $\boldsymbol{s} \in \mathbb{R}^N$ be a vector of N M-Estimators where L_i is the corresponding loss-function for M-Estimator s_i $(1 \leq i \leq N)$. Let $L = \sum_{i=1}^{N} L_i$. Then, the Robust Covariance Matrix of $\boldsymbol{s}$ is given by

$$\tilde{\boldsymbol{\Sigma}}(\boldsymbol{s}) = \boldsymbol{H}_L^{-1}(\boldsymbol{s})\boldsymbol{V}_L(\boldsymbol{s})\boldsymbol{H}_L^{-1}(\boldsymbol{s})$$

where $\boldsymbol{H}_L$ is the Hessian matrix of L and $\boldsymbol{V}_L$ is the sample covariance of the gradient ∇L taken over y_i^1, $1 \leq i \leq m$. That is $\boldsymbol{V}_L(\boldsymbol{s}) = \text{cov}(\boldsymbol{M}_L(\boldsymbol{s}))$ where

$$\boldsymbol{M}_L(\boldsymbol{s}) = \begin{pmatrix} — & (\nabla L(\boldsymbol{s}))|_{y_1^1} & — \\ — & (\nabla L(\boldsymbol{s}))|_{y_2^1} & — \\ \vdots & \vdots & \vdots \\ — & (\nabla L(\boldsymbol{s}))|_{y_m^1} & — \end{pmatrix}.$$

In Definition 3.5, $(\nabla L(\boldsymbol{s}))|_{y_i^1}$ refers to the gradient $\nabla L(\boldsymbol{s})$ evaluated at the point of the trajectory y_i^1.

All the statistics chosen in the Zero-Inflated Gamma example are M-Estimators. Therefore, the Robust Covariance Matrix can be applied to the Zero-Inflated Gamma example.

3.4 Example: Zero-Inflated Gamma Continued

The setup is identical to before. Start with a ZIG(α, β, p) random variable where $\alpha^* = 3$, $\beta^* = 1$ and $p^* = 0.3$, obtain the observed trajectory $\boldsymbol{y}^0$ and calculate the vector of observed summary statistics $\boldsymbol{s}^0 = S(\boldsymbol{y}^0)$. Now, construct the likelihood function. First,

a single sample, denoted $\boldsymbol{y}^1$, is generated. Take

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \boldsymbol{s}^1 = S(\boldsymbol{y}^1) \tag{7}$$

and

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \approx \tilde{\boldsymbol{\Sigma}}(\boldsymbol{s}^1) = \boldsymbol{H}_L^{-1}(\boldsymbol{s}^1)\boldsymbol{V}_L(\boldsymbol{s}^1)\boldsymbol{H}_L^{-1}(\boldsymbol{s}^1) \tag{8}$$

where L is the loss function corresponding to $S(\cdot)$, the vector of summary statistics. To find L, each statistic must be written as a solution to the minimisation of a loss function. Recall the statistics are

$$S(\boldsymbol{y}^i) = (\bar{x}(\boldsymbol{y}^i), \mathrm{sd}(\boldsymbol{y}^i), \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\{y_j^i = 0\}). \tag{9}$$

Since sample mean $\bar{x}(\boldsymbol{y}^i)$ and standard deviation $\mathrm{sd}(\boldsymbol{y}^i)$ are Maximum Likelihood Estimators, they are M-Estimators as per Proposition 3.3. In particular, they are the solution $(\xi_1, \xi_2) = (\bar{x}(\boldsymbol{y}^i), \mathrm{sd}(\boldsymbol{y}^i))$ to the minimisation of

$$L_{1,2}(\xi_1, \xi_2) = -\sum_{i=1}^{m} \log(f(y_i^1 \mid \xi_1, \xi_2^2)) \tag{10}$$

where f is the density of a Normal distribution with mean ξ_1 and standard deviation ξ_2. For the proportion of zeros statistic, a quadratic loss function where $\xi_3 = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}(y_j^i = 0)$ is the solution to the minimisation of

$$L_3(\xi_3) = \frac{1}{2}\sum_{i=1}^{m} (\xi_3 - \mathbb{I}\{y_i^1 = 0\})^2 \tag{11}$$

can be used. Therefore the loss function is

$$L(\xi_1, \xi_2, \xi_3) = -\sum_{i=1}^{m} \log(f(y_i^1 \mid \xi_1, \xi_2^2)) + \frac{1}{2}\sum_{i=1}^{m} (\xi_3 - \mathbb{I}\{y_i^1 = 0\})^2. \tag{12}$$

Now the Hessian and the covariance of the gradient of L, evaluated at $(\xi_1, \xi_2, \xi_3) = \boldsymbol{s}^1$, are needed. By the chain rule, the gradient of L for a single y_i^1 is

$$(\nabla L(\xi_1, \xi_2, \xi_3))|_{y_i^1} = (-\frac{y_i^1 - \xi_1}{\xi_2^2}, \frac{\xi_2^2 - (y_i^1 - \xi_1)^2}{\xi_2^3}, \xi_3 - \mathbb{I}\{y_i^1 = 0\}). \tag{13}$$

Then, taking the covariance over all observations yields

$$\boldsymbol{V}_L(\xi_1, \xi_2, \xi_3) = \text{cov} \begin{pmatrix} — & (\nabla L(\xi_1, \xi_2, \xi_3))|_{y_1^0} & — \\ — & (\nabla L(\xi_1, \xi_2, \xi_3))|_{y_2^0} & — \\ \vdots & \vdots & \vdots \\ — & (\nabla L(\xi_1, \xi_2, \xi_3))|_{y_m^0} & — \end{pmatrix}. \tag{14}$$

The Hessian can be computed numerically. To compute the Synthetic Likelihood:

1. Simulate the trajectory $\boldsymbol{y}^1$
2. Compute $\hat{\boldsymbol{\mu}} = \boldsymbol{s}^1 = S(\boldsymbol{y}^1)$
3. Plug in $\boldsymbol{s}^1$ for (ξ_1, ξ_2, ξ_3) to estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \approx \tilde{\boldsymbol{\Sigma}}(\boldsymbol{s}^1) = \boldsymbol{H}_L^{-1}(\boldsymbol{s}^1)\boldsymbol{V}_L(\boldsymbol{s}^1)\boldsymbol{H}_L^{-1}(\boldsymbol{s}^1)$
4. Obtain the Synthetic Likelihood $\hat{p}(\boldsymbol{s}^0|\boldsymbol{\theta}) = \varphi(\boldsymbol{s}^0|\boldsymbol{s}^1, \tilde{\boldsymbol{\Sigma}}(\boldsymbol{s}^1))$

Again, this likelihood function can be passed to the 'ML' optimizer from *Synlik* (Matteo Fasiolo and Simon N. Wood 2014) to produce the estimate $\hat{\boldsymbol{\theta}}$. Convergence to the true parameters is still achieved and shown in Figure 6. Noteably, this happens much faster (time wise) since 100× fewer samples are used.

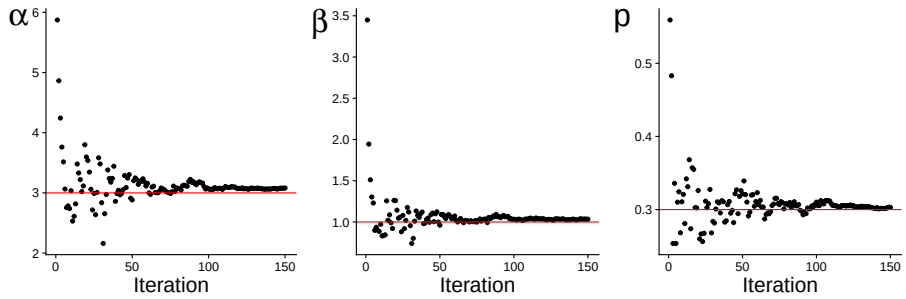

Figure 6: Convergence to True Parameters (ZIG with RCM)

The next section will propose a heuristic for determining whether a set of statistics are likely to provide a good Synthetic Likelihood Estimate of the parameters.

3.5 What is a Good Statistic?

A good statistic must have the following two characteristics:

- High correlation with model parameters,

- Low correlation with other statistics.

The first criterion ensures the statistics provide meaningful information about the model parameters. The second criterion is also important as too much mutual correlation could result in singular covariance matrices. Thus, correlation diagrams, as shown in Figure 7, can be produced to build an idea for how a set of statistics may perform. These are particularly useful since it is far more computationally efficient to compute correlations than it is to perform Synthetic Likelihood Estimation.

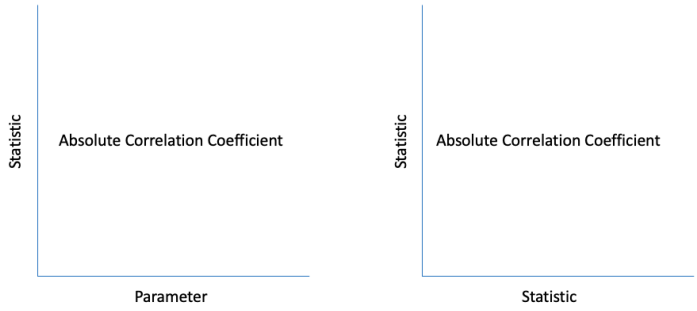

Figure 7: Correlation Diagrams

Concretely, to produce these diagrams, choose sample parameters $\boldsymbol{\theta}$ and covariance matrix Σ. Then, several sets of parameters from $\mathcal{N}(\boldsymbol{\theta}, \Sigma)$, a multivariate Normal distribution, are sampled. For each parameter, an observed trajectory is produced and its summary statistics evaluated. The correlations between mutual statistics and between statistics and parameters can then be computed. Algorithm 2 defines this process precisely. For notation, the i'th row of matrix X is denoted by X_{i*} and the j'th column by X_{*j}.

Algorithm 2 Correlation Diagrams

Let $\boldsymbol{\theta}$ and Σ be the parameters of the multivariate Normal distribution from which model parameters will be drawn
Let $f(\cdot)$ be a function that computes a random trajectory for some given parameters
Let $S(\cdot)$ be a function that computes the statistics for a given trajectory
Let n be the number of parameters to sample and m be the number of statistics computed by $S(\cdot)$

procedure STAT_COR($\boldsymbol{\theta}, \Sigma, f, S, n, m$)
    $P \leftarrow n$ samples from $\mathcal{N}(\boldsymbol{\theta}, \Sigma)$
    $S \leftarrow$ Empty matrix of dimensions $n \times m$
    for $i \leftarrow 1 : n$ **do**
        $\boldsymbol{t} \leftarrow f(P_{i*})$
        $S_{i*} \leftarrow S(\boldsymbol{t})$
    end for
    $C \leftarrow$ Empty matrix of dimensions $m \times \dim(\boldsymbol{\theta})$
    for $i \leftarrow 1 : \dim(\boldsymbol{\theta})$ **do**
        for $j \leftarrow 1 : m$ **do**
            $C_{ji} = |\text{cor}(P_{*i}, S_{*j})|$
        end for
    end for
    $D \leftarrow |\text{cor}(S)|$
    return C, D
end procedure

Note: C is the matrix of correlations between statistics and parameters whereas D is the matrix of correlations between statistics mutually.

As a note from experience, it is best to first obtain a set of statistics that have strong correlations with all model parameters. Then, if Synthetic Likelihood Estimation fails, remove the most correlated statistics and repeat.

These diagrams can be computed for the Zero-Inflated Gamma Example. They are displayed in Figure 8. One can see each parameter is correlated with at least one of the statistics. There is correlation between the mean and standard deviation statistics but this is not enough to upset the Synthetic Likelihood Estimation.

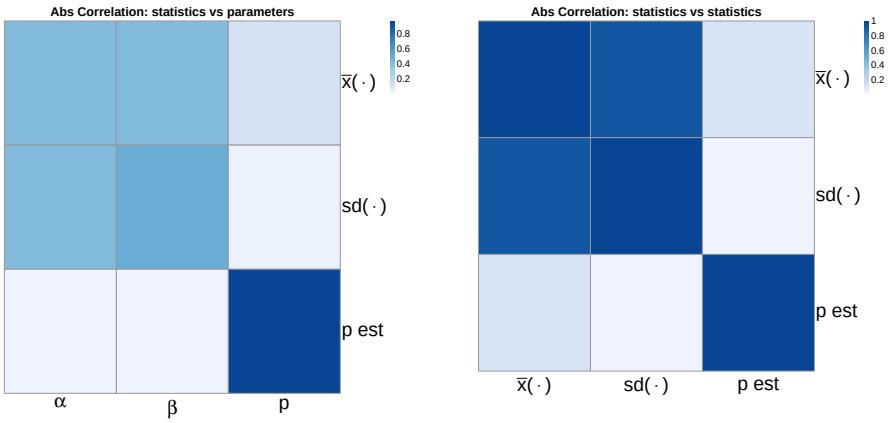

Figure 8: ZIG Example Correlation Diagrams

The next section will take a closer look at the 'ML' optimiser from *Synlik* (Matteo Fasiolo and Simon N. Wood 2014).

3.6 ML Optimiser

The optimiser works as follows. It takes in initial values for $\boldsymbol{\theta}$ and an initial covariance Σ. Then it samples n parameters from $\mathcal{N}(\boldsymbol{\theta}, \Sigma)$, a multivariate Normal distribution. For each sample, the likelihood function is evaluated. It then chooses the next $\boldsymbol{\theta}$ as a linear combination of the samples, weighted by their Synthetic Likelihood. The next Σ is chosen to be $\alpha\Sigma$ where $\alpha \in (0, 1)$ is a constant; this ensures convergence. The algorithm is precisely defined in Algorithm 3.

Algorithm 3 ML Optimiser

Let f be a function that computes the Synthetic Likelihood for a set of parameters
Let $\boldsymbol{\theta}_0$ be the initial parameters
Let Σ be the initial covariance
Let n be the number of parameters to sample
Let N be the number of iterations

procedure ML($f, \boldsymbol{\theta}_0, \Sigma, n, N, \alpha$)
    $\Theta \leftarrow$ Empty matrix of dimensions $N + 1 \times \dim(\boldsymbol{\theta}_0)$
    $\Theta_{1*} \leftarrow \boldsymbol{\theta}_0^T$
    for i in 1:N **do**
        $P \leftarrow n$ samples from $\mathcal{N}(\Theta_{i*}, \Sigma)$
        $\boldsymbol{l} \leftarrow (f(P_{j*}))_{j=1}^{n}$
        Discard any undefined likelihoods from $\boldsymbol{l}$ and the corresponding rows of P
        Let n_2 be the number of well-defined likelihoods
        Define $g(x) = e^{x-m}$ where $m = \max\{l_1, \ldots, l_{n_2}\}$
        $\boldsymbol{u} \leftarrow (g(l_j))_{j=1}^{n_2}$
        $\boldsymbol{w} \leftarrow (\sum_{j=1}^{n_2} u_j)^{-1}\boldsymbol{u}$
        Obtain X_1 by subtracting Θ_{i*} from each row of P
        Obtain X_2 by multiplying each row of X_1 by its corresponding weight in $\boldsymbol{w}$
        Obtain $\boldsymbol{v}$ by summing each column of X_2
        $\boldsymbol{\theta}_1 \leftarrow (\Theta_{ij} - v_j)_{j=1}^{\dim(\boldsymbol{\theta}_0)}$
        $\Theta_{i+1,*} \leftarrow \boldsymbol{\theta}_1^T$
    end for
    return Θ
end procedure

The next section will prove some additional important M-Estimators.

3.7 M-Estimators

In Section 3.3, it was shown that Maximum Likelihood Estimators are M-Estimators. In preparation for fitting Model $\mathcal{A}$, it will also be shown that OLS regression estimators and the sample mean under a quadratic loss function are also M-Estimators.

First, that OLS regression estimators are M-Estimators.

Proposition 3.6. *Suppose there is data* $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \boldsymbol{y}$. *Consider the regression* $\boldsymbol{y} =$

$X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *where*

$$X = \begin{pmatrix} 1 & — & \boldsymbol{x}_1^T & — \\ 1 & — & \boldsymbol{x}_2^T & — \\ \vdots & \vdots & \vdots & \vdots \\ 1 & — & \boldsymbol{x}_n^T & — \end{pmatrix}. \tag{15}$$

Then, the OLS estimators $\hat{\boldsymbol{\beta}}$ are M-Estimators.

Proof. As per OLS estimation, obtain $\hat{\boldsymbol{\beta}}$ by minimising the sum of squared residuals (Hayashi 2000, p. 15). That is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \epsilon_i^2 = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - X_{i*}\boldsymbol{\beta})^2. \tag{16}$$

Thus, taking $\rho(X_{i*}, y_i, \boldsymbol{\beta}) = (y_i - X_{i*}\boldsymbol{\beta})^2$ satisfies what is needed for $\hat{\boldsymbol{\beta}}$ to be an M-Estimator. □

Second, that the sample mean with the quadratic loss function is an M-Estimator.

Proposition 3.7. *Consider data $\boldsymbol{x} \in \mathbb{R}^n$. Then the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ minimises the quadratic loss $L(\xi) = \frac{1}{2}\sum_{i=1}^{n}(x_i - \xi)^2$ and is thus an M-Estimator.*

Proof. Taking the first derivative with respect to ξ yields

$$\frac{\mathrm{d}L}{\mathrm{d}\xi}(\xi) = \sum_{i=1}^{n}(\xi - x_i) = n\xi - \sum_{i=1}^{n} x_i. \tag{17}$$

So L has a unique stationary point at $\xi = \bar{x}$. Taking the second derivative reveals

$$\frac{\mathrm{d}^2 L}{\mathrm{d}\xi^2}(\xi) = n > 0 \tag{18}$$

so $\xi = \bar{x}$ is the unique minimum of L and thus sample mean $\bar{x}$ with the quadratic loss function is an M-Estimator. □

Finally, before fitting Model $\mathcal{A}$, the asymptotic normality of M-Estimators will be presented.

3.8 Asymptotic Normality of M-Estimators

For the purposes of this section, $\to_{\mathcal{D}}$ will denote convergence in distribution and $\to_{\mathbb{P}}$ will denote convergence in probability.

The goal will be to show that an M-Estimator is asymptotically normal in the following sense. Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables. Also suppose $(\hat{\theta}_n)$ is a sequence of $\mathcal{M}$-estimators. That is

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \rho(X_i, \theta) \tag{19}$$

for some function ρ. Then, under certain conditions and using 'sloppy' asymptotics $\hat{\theta}_n \approx \mathcal{N}(\theta^*, \tilde{\Sigma})$ where

$$\theta^* = \operatorname*{argmin}_{\theta} \mathbb{E}\rho(X_1, \theta) \tag{20}$$

and $\tilde{\Sigma}$ is the Robust Covariance Matrix. This is an important theorem as it justifies the use of the Robust Covariance Matrix and the multivariate Normal assumption for Synthetic Likelihood Estimation.

First, a sketch proof will be presented, following the approach from Geyer (2013). Then, a more rigorous theorem from Pollard (1985) will be stated but not proved.

Define $\lambda(\theta) = \mathbb{E}\rho(X_1, \theta)$. Assume λ achieves its minimum at some point θ^* on the interior of the parameter space. So $\lambda'(\theta^*) = 0$. Next, write $\rho_n(\theta) = \sum_{i=1}^{n} \rho(X_i, \theta)$. Now, define

$$V_n(\theta) = \operatorname{var}(\rho_n'(\theta)) \tag{21}$$

and

$$J_n(\theta) = -\mathbb{E}\rho_n''(\theta). \tag{22}$$

By properties of the variance and expectation, $V_n(\theta) = nV_1(\theta)$ where $V_1(\theta) = \operatorname{var}(\rho'(X_1, \theta))$ and $J_n(\theta) = nJ_1(\theta)$ where $J_1(\theta) = \mathbb{E}\rho''(X_1, \theta)$. Now, under some technical conditions and the law of large numbers

$$-\frac{1}{n}\rho_n''(\theta^*) \to_{\mathbb{P}} J_1(\theta^*) \tag{23}$$

and using the central limit theorem,

$$\frac{1}{\sqrt{n}}\rho_n'(\theta^*) \to_{\mathcal{D}} \mathcal{N}(0, V_1(\theta^*)). \tag{24}$$

Considering the Taylor expansion of ρ'_n around θ^* (this is justified for large n since it turns out $\hat{\theta}_n$ is consistent) reveals

$$\rho'_n(\hat{\theta}_n) \approx \rho'_n(\theta^*) + \rho''_n(\theta^*)(\hat{\theta}_n - \theta^*). \tag{25}$$

Since $\hat{\theta}_n$ minimises $\rho_n(\theta)$, it must be that $\rho'(\hat{\theta}_n) = 0$. Thus, setting the right hand side of Equation 25 to zero, one obtains

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \approx \frac{-\frac{1}{\sqrt{n}}\rho'_n(\theta^*)}{\frac{1}{n}\rho''_n(\theta^*)}. \tag{26}$$

Thus, an application of Slutsky's theorem gives

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \to_{\mathcal{D}} \mathcal{N}(0, J_1(\theta^*)^{-1}V_1(\theta^*)J_1(\theta^*)^{-1}). \tag{27}$$

which is *almost* what was desired. To finish it off, a second use of Slutsky's theorem with the consistency of $\hat{\theta}_n$ and applying 'sloppy' asymptotics one obtains

$$\hat{\theta}_n \approx \mathcal{N}(\theta^*, \hat{J}_n(\hat{\theta}_n)^{-1}\hat{V}_n(\hat{\theta}_n)\hat{J}_n(\hat{\theta}_n)^{-1}) \tag{28}$$

where $\hat{J}_n(\theta) = -\rho''_n(\theta)$, the observed Fisher information and $\hat{V}_n(\theta) = \sum_{i=1}^n \rho'_n(\theta)^2$. This sketch proof works almost identically if θ is a vector instead of a scalar. In this case, $\hat{J}_n$ is the Hessian and $\hat{V}_n$ is the covariance of the gradient of ρ_n. Thus, $\hat{J}_n(\hat{\theta}_n)^{-1}\hat{V}_n(\hat{\theta}_n)\hat{J}_n(\hat{\theta}_n)^{-1}$ is *precisely* the Robust Covariance Matrix.

Now for a more rigorous theorem from Pollard (1985). First, stochastic differentibility will be defined. Then, the theorem itself will be stated.

Definition 3.8. Let ρ be a function and θ^* a point on the interior of the domain of ρ. Stochastic differentiability of ρ at θ^* means there exists a linear approximation of ρ near θ^* with remainder term $r(x, \theta)$ that is small 'in an average sense' compared with $|\theta - \theta^*|$. Concretely, this means one can write

$$\rho(x, \theta) = \rho(x, \theta^*) + (\theta - \theta^*)^T\Delta(x) + |\theta - \theta^*|r(x, \theta) \tag{29}$$

where Δ denotes a vector of functions that depend only on x.

A more rigorous notation of what it means for $r(x, \theta)$ to be small 'in an average sense' compared with $|\theta - \theta^*|$ is given in Pollard (1985) as the Stochastic Differentiability Condition. Now for the theorem.

Theorem 3.9. *Let $X_1, X_2, \ldots$ be independent observations from a distribution P. Write $\mathcal{P}f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ for some function f. Let $(\hat{\theta}_n)$ be the sequence of estimators defined as in Equation 19. Suppose*

1. *$G(\theta) := \mathbb{E}\rho(X_1, \theta)$ has invertible second derivative $-J$ at its maximising value θ^*;*
2. *θ^* is an interior point of the parameter space;*
3. *$\hat{\theta}_n \to_{\mathbb{P}} \theta^*$;*
4. *ρ is stochastically differentiable as per Equation 29;*
5. *Vector Δ in Equation 29 satisfies $\mathcal{P}\Delta = 0$ and $\mathcal{P}|\Delta|^2 < \infty$.*

Then $\sqrt{n}(\hat{\theta}_n - \theta^*) \to_{\mathcal{D}} \mathcal{N}(0, J^{-1}\mathcal{P}(\Delta\Delta^T)J^{-1})$.

The result of this theorem matches what was found with the sketch proof but is more rigorous. The conclusion is that using the Robust Covariance Matrix as the covariance of a multivariate Normal distribution makes sense and the multivariate Normal assumption is justified.

The next section will tackle the fitting of Model $\mathcal{A}$.

4 Fitting

This section will fit Model $\mathcal{A}$ using Synthetic Likelihood Estimation. Since Model $\mathcal{A}$ is quite complicated, this section will start by looking at a simplified model where there is no regime switching. Fitting this simplified model will provide the building blocks to fit Model $\mathcal{A}$.

To start with, synthetic data will be used instead of real data. A sample can be drawn using fitted values for the UK power data (Table 1) from Huisman & Mahieu (2003). This sample will be used as the observed trajectory which will then be used to fit the model. This allows for a comparison between the fitted parameters and the original parameters to check the statistics properly explain the dynamics of the model.

In Section 4.2.3, historic data from Nordpool will be used to fit the model.

4.1 Fitting a Simplified Model (Model $\mathcal{B}$)

The cleanest way to simplify Model $\mathcal{A}$ is to forget the regime-switching part and stay in regime 0 for all time. This means the simplified model is specified as follows. Model

the natural log of the spot price $s(t)$ as the sum of the deterministic component $f(t) = \mu_0 + \beta_1 D_1(t) + \beta_2 D_2(t)$ and the stochastic component $x(t)$. However, the stochastic component is now modelled as $x(1) = 0$ and

$$\mathrm{d}x(t) = x(t) - x(t-1) = -\alpha_0 x(t-1) + \sigma_0 \epsilon_t \tag{30}$$

for $t > 1$. Constrained parameters α_0 and σ_0 still need to be replaced with unconstrained parameters τ_{α_0} and τ_{σ_0} as per Section 2.2. Call the above model, Model $\mathcal{B}$.

4.1.1 Choice of Statistics

Now, the statistics need to be chosen. For μ_0 and σ_0, include the Maximum Likelihood Estimates from a Normal distribution. Denote the sample mean as $\bar{x}(\cdot)$ and standard deviation as $\mathrm{sd}(\cdot)$. Additionally, β_1 and β_2 represent the price increment on a Saturday and a Sunday respectively. Thus, to infer these parameters, include the mean price on a Saturday, denoted as $\bar{x}_5(\cdot)$, and on a Sunday, denoted as $\bar{x}_6(\cdot)$. Notice, Monday is considered day 0, so Saturday is day 5 and Sunday is day 6 in this notation. Finally, statistics are needed for α_0, the mean-reversion parameter. There are a few statistics that may be useful. First, mean-reversion is to do with the relationship between $x(t)$ and $x(t-1)$. Thus, it would make sense to include autocorrelation coefficients. However, these are not M-estimators. Taking a closer look at the dynamics of Model $\mathcal{B}$ reveals a similar option - autoregression coefficients. Expanding the Model $\mathcal{B}$ yields

$$\begin{aligned}
s(t) &= f(t) + x(t) && (31)\\
&= f(t) + x(t-1) + \mathrm{d}x(t) && (32)\\
&= f(t) + x(t-1) - \alpha_0 x(t-1) + \sigma_0 \epsilon_t && (33)\\
&= f(t) + (1-\alpha_0) x(t-1) + \sigma_0 \epsilon_t && (34)\\
&= f(t) + (1-\alpha_0)(s(t-1) - f(t-1)) + \sigma_0 \epsilon_t && (35)\\
&= f(t) + (1-\alpha_0) f(t-1) + (1-\alpha_0) s(t-1) + \sigma_0 \epsilon_t. && (36)
\end{aligned}$$

The expanded dynamics show that an autoregression of the form

$$s(t) \approx \xi_0 + \xi_1 s(t-1) \tag{37}$$

is appropriate, with the gradient parameter, ξ_1, of particular interest. Since OLS estimators, denoted $\xi_0(\cdot)$ and $\xi_1(\cdot)$ respectively, are M-Estimators as per proposition 3.6,

all the statistics are M-Estimators. Thus, Synthetic Likelihood Estimation can be compared with and without using the Robust Covariance Matrix. Before proceeding, the correlation diagrams defined in Section 3.5 will be computed.

These are shown in Figure 9. Each parameter is reasonably correlated with one of the statistics and the statistics themselves are not too correlated. There are two classes of notable correlation between the statistics. The first is between the mean spot price, the mean spot price on Saturday and the mean spot price on Sunday. This could be fixed by subtracting the mean from the mean on Saturday and the mean Sunday, but the resulting statistics would no longer be M-Estimators. The other area of notable correlation is between ξ_1 and ξ_2, the autoregression intercept and gradient. This correlation is expected since both parameters are strongly correlated with α_0. Both statistics need to be included in order to use the Robust Covariance Matrix.

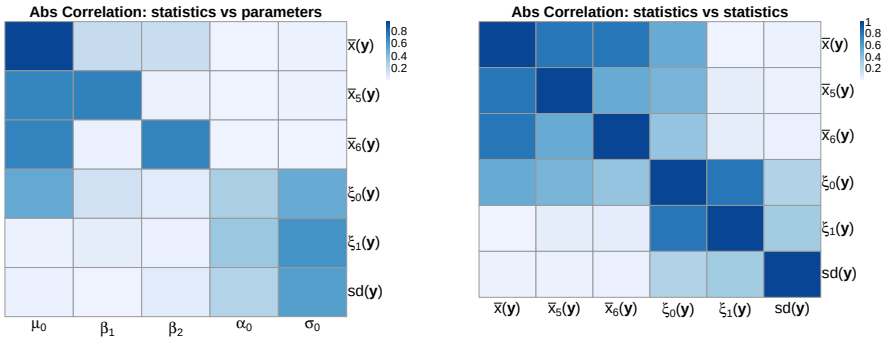

Figure 9: Model $\mathcal{B}$ Correlation Diagrams

In order to compute the Robust Covariance Matrix, a loss function is needed. Let m be the observation length and $\boldsymbol{y}^0$ the observed trajectory. A sample trajectory $\boldsymbol{y}^1$ is also needed which the statistics will be evaluated along. For $\bar{x}(\cdot)$ and $\mathrm{sd}(\cdot)$, the negative log-likelihood of a Normal distribution will suffice. Write

$$L_{1,2} = -\sum_{i=1}^{m} \log f(y_i^1 | s_1, s_2^2) \tag{38}$$

where f is the density of a Normal random variable with mean $s_1 = \bar{x}(\boldsymbol{y}^1)$ and standard deviation $s_2 = \mathrm{sd}(\boldsymbol{y}^1)$.

Now, for $\bar{x}_5(\cdot)$ and $\bar{x}_6(\cdot)$, the mean price on a Saturday and a Sunday respectively. Here,

a quadratic loss function will be used. That is,

$$L_3 = \frac{1}{2}\sum_{i=1}^{m}(s_3 - y_i^1)^2 \mathbb{I}\{i \equiv 5 \pmod 7\} \tag{39}$$

and

$$L_4 = \frac{1}{2}\sum_{i=1}^{m}(s_4 - y_i^1)^2 \mathbb{I}\{i \equiv 6 \pmod 7\} \tag{40}$$

where $s_3 = \bar{x}_5(\boldsymbol{y}^1)$ and $s_4 = \bar{x}_6(\boldsymbol{y}^1)$.

Finally, for the OLS regression estimators, the sum of squared residuals loss function will be used. Write

$$L_{5,6} = \sum_{i=2}^{m}[y_i^1 - (s_5 + s_6 y_{i-1}^1)]^2 \tag{41}$$

where $s_5 = \xi_0(\boldsymbol{y}^1)$ and $s_6 = \xi_1(\boldsymbol{y}^1)$.

These individual loss functions can now be summed to obtain a loss function for the vector of statistics $\boldsymbol{s} = (s_1, \ldots, s_6)$. Concretely,

$$\begin{aligned} L = &-\sum_{i=1}^{m}\log f(y_i^1|s_1, s_2^2) + \frac{1}{2}\sum_{i=1}^{m}(s_3 - y_i^1)^2\mathbb{I}\{i \equiv 5 \pmod 7\} \\ &+ \frac{1}{2}\sum_{i=1}^{m}(s_4 - y_i^1)^2\mathbb{I}\{i \equiv 6 \pmod 7\} + \sum_{i=2}^{m}[y_i^1 - (s_5 + s_6 y_{i-1}^1)]^2. \end{aligned} \tag{42}$$

The next step is to compute the gradient and Hessian of L. The gradient turns out to be

$$(\nabla L(\boldsymbol{s}))|_{y_i} = \begin{pmatrix} -\frac{y_i^1 - s_1}{s_2^2} \\ \frac{s_2^2 - (y_i^1 - s_1)^2}{s_2^3} \\ (s_3 - y_i^1)\mathbb{I}\{i \equiv 5 \pmod 7\} \\ (s_4 - y_i^1)\mathbb{I}\{i \equiv 6 \pmod 7\} \\ 2(-y_i^1 + y_{i-1}^1 s_6 + s_5)\mathbb{I}\{i > 1\} \\ 2y_{i-1}^1(-y_i^1 + y_{i-1}^1 s_6 + s_5)\mathbb{I}\{i > 1\} \end{pmatrix}. \tag{43}$$

To achieve the greatest computational benefit from using the Robust Covariance Matrix, the Hessian will also be computed analytically. This is just a matter of further differentiating the gradient. It is also possible to compute the Hessian numerically; this will be considered for the performance section (Section 4.1.3). For the analytic Hessian,

obtain

$$\boldsymbol{H}_L(\boldsymbol{s}) = \sum_{i=1}^{m} \begin{pmatrix} \frac{1}{s_2^2} & \frac{2(y_i^1 - s_1)}{s_2^3} & 0 & 0 & 0 & 0 \\ \frac{2(y_i^1 - s_1)}{s_2^3} & -\frac{1}{s_2^2} + \frac{3(y_i^1 - s_1)^2}{s_2^4} & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbb{I}\{i \equiv 5 \pmod 7\} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbb{I}\{i \equiv 6 \pmod 7\} & 0 & 0 \\ 0 & 0 & 0 & 0 & 2\mathbb{I}\{i > 1\} & 2y_{i-1}^1 \mathbb{I}\{i > 1\} \\ 0 & 0 & 0 & 0 & 2y_{i-1}^1 \mathbb{I}\{i > 1\} & 2(y_{i-1}^1)^2 \mathbb{I}\{i > 1\} \end{pmatrix}. \quad (44)$$

This can be slightly simplified by bringing the sum into the matrix yielding

$$\boldsymbol{H}_L(\boldsymbol{s}) = \begin{pmatrix} \frac{m}{s_2^2} & \sum_{i=1}^m \frac{2(y_i^1 - s_1)}{s_2^3} & 0 & 0 & 0 & 0 \\ \sum_{i=1}^m \frac{2(y_i^1 - s_1)}{s_2^3} & -\frac{m}{s_2^2} + \sum_{i=1}^m \frac{3(y_i^1 - s_1)^2}{s_2^4} & 0 & 0 & 0 & 0 \\ 0 & 0 & \lfloor \frac{m+2}{7} \rfloor & 0 & 0 & 0 \\ 0 & 0 & 0 & \lfloor \frac{m+1}{7} \rfloor & 0 & 0 \\ 0 & 0 & 0 & 0 & 2(m-1) & 2\sum_{i=2}^m y_{i-1}^1 \\ 0 & 0 & 0 & 0 & 2\sum_{i=2}^m y_{i-1}^1 & 2\sum_{i=2}^m (y_{i-1}^1)^2 \end{pmatrix}. \quad (45)$$

Now, all the pieces to perform Synthetic Likelihood Estimation, with and without the Robust Covariance Matrix, are in place. All that is left is to check the optimisation converges to the true parameters and to compare the computational performance with and without using the Robust Covariance Matrix.

4.1.2 Checking Convergence

A simulation study can be performed to confirm that the optimisation converges to the true parameters. A sample is generated from the true parameters. This sample will be used as the observed trajectory. Synthetic Likelihood Estimation is then performed, with and without the Robust Covariance Matrix to obtain fitted values. This process is repeated and the fitted values can then be compared to the true values.

In this case, 50 simulations were completed. When not using the Robust Covariance Matrix, 100 samples were drawn to compute each Synthetic Likelihood.

In Figure 10, box plots can be found for each parameter. Here, the true parameter value is subtracted from each parameter. Thus, a good fit would show clustering around zero, with small variance, for each parameter. The Root Mean Squared Errors (RMSE) are also of interest and can be found in Table 2.

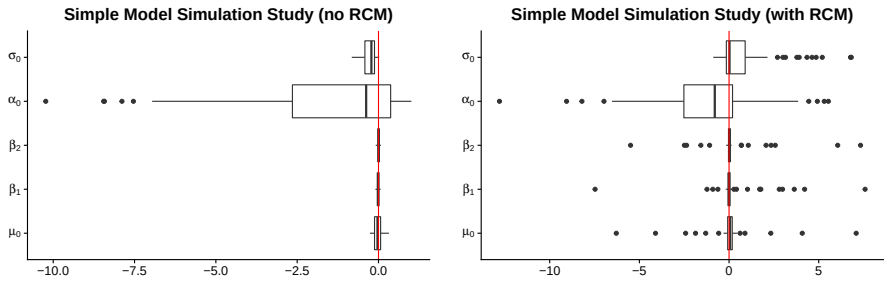

Figure 10: Model $\mathcal{B}$ Simulation Study Box Plots

Table 2: Root Mean Squared Errors for Model $\mathcal{B}$

Parameter	RMSE without RCM (3 s.f.)	RMSE with RCM (3 s.f.)
μ_0	0.126	1.69
β_1	0.0366	1.85
β_2	0.0352	1.76
τ_{α_0}	3.42	3.79
τ_{σ_0}	0.353	2.23

The results show that Synthetic Likelihood Estimation, both with and without the Robust Covariance Matrix can achieve convergence to the true parameters. Notably, using the Robust Covariance Matrix increases the RMSE and the number of outliers, this is to be expected since it is a noisy estimator. However, as will be seen when looking at the performance, the Robust Covariance Matrix runs much more quickly. Thus, the observation length can be increased far past what would be computationally reasonable when not using the Robust Covariance Matrix. This itself can significantly should reduce the RMSE due to the asymptotic normality of the M-Estimator (Thereom 3.9).

4.1.3 Performance

Moving on to performance. This can be analysed by repeatedly timing how long it takes to evaluate a Synthetic Likelihood with and without the Robust Covariance Matrix. In this case, the Synthetic Likelihood of the true parameters (Table 1) were evaluated using the R library *microbenchmark*. In Figure 11, violin plots of the evaluation times are shown. The numerical Robust Covariance Matrix is evaluated using the 'fdHess' function from the R library *nlme*. It is clear that the Robust Covariance Matrix pro-

vides a significant performance boost. Specifically, the median time when not using the Robust Covariance Matrix was 6ms, about 37× longer than 0.16ms, the median time when using the Analytic Robust Covariance Matrix. Additionally, not using the Robust Covariance Matrix takes about 7× longer than 0.87ms, the median time when using the Numerical Robust Covariance Matrix. Thus, including the Robust Covariance Matrix with a numerical Hessian estimate significantly improves computational performance. However, the bulk of the computational performance improvement comes when using an analytic Hessian.

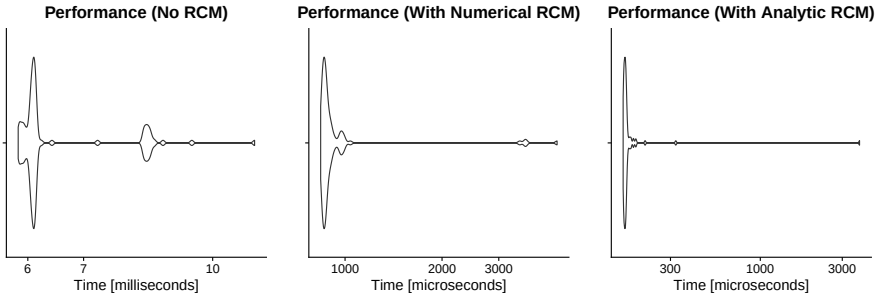

Figure 11: Comparative Performance of Robust Covariance Matrix

This section has demonstrated how using the Robust Covariance Matrix can be used within Synthetic Likelihood Estimation to provide significant computational benefit.

The next section will address fitting Model $\mathcal{A}$.

4.2 Fitting Model $\mathcal{A}$

4.2.1 Choice of Statistics

Now that Model $\mathcal{B}$ has been successfully fit, it is time to return to the Model $\mathcal{A}$. Statistics need to be chosen that capture the nature of all parameters. For clarity, the parameters will be split up into the following groups. The weekly shape parameters: μ_0, β_1 and β_2. The volatility parameters: σ_0, σ_1 and σ_{-1}. The mean reversion parameters: α_0 and α_{-1}. Finally, the spike parameters: μ_1 and p.

To fit the Model $\mathcal{A}$, statistics will need to be chosen that are not M-estimators. Future projects may try to find equivalent M-estimators.

Before diving into the individual parameters, as shown in Wood (2010), including statis-

tics that quantify the 'shape of the marginal distribution' can be useful. These can be obtained as the coefficients of the polynomial regression between the ordered differences of the simulated trajectory and the observed trajectory. Ordered differences means the differences $y_{i+1} - y_i$ ordered from smallest to largest. This proves to be a good way to ensure the trajectories produced from fitted parameters follow similar distributions to the distribution of the observed trajectory. This project will use cubic regression and denote the coefficients $\gamma_0(\cdot), \gamma_1(\cdot), \gamma_2(\cdot), \gamma_3(\cdot)$.

For the shape parameters, similar statistics to those chosen for Model $\mathcal{B}$ can be used. For example, for μ_0, the sample mean denoted $\bar{x}(\cdot)$, can be used. For β_1, the sample mean taken over Saturdays minus the sample mean can be used. This can be written as $(\bar{x}_5 - \bar{x})(\cdot)$. Subtracting the sample mean reduces the correlation between this statistic and the sample mean which will improve the Synthetic Likelihood Estimate. Similarly, β_2 can be inferred from $(\bar{x}_6 - \bar{x})(\cdot)$.

There are several statistics that provide useful information about the volatility parameters. Most obviously, the sample standard deviation, denoted $\text{sd}(\cdot)$, and the interquartile range, denoted $\text{IQR}(\cdot)$. However, the maximum of the absolute values of the trajectory will also exhibit some correlation with these parameters. Additionally, first and second order differences may also be of interest. That is, trajectory $\boldsymbol{d}^1$ can be formed by $d_i^1 = y_{i+1} - y_i$ and trajectory $\boldsymbol{d}^2$ by $d_i^2 = d_{i+1} - d_i$. Then, summary statistics for the distributions of $\boldsymbol{d}^1$ and $\boldsymbol{d}^2$ can be considered. Concretely, statistics such as the sample mean, standard deviation, skew and kurtosis can be calculated for $\boldsymbol{d}^1$ and $\boldsymbol{d}^2$. The sample skew shall be denoted as $\text{sk}(\cdot)$ and sample kurtosis as $\kappa(\cdot)$.

Another statistic that may be of interest for the volatility parameters is the maximum absolute difference, denoted $\max|\cdot|$. This statistic can be calculated for $\boldsymbol{y}, \boldsymbol{d}^1$ and $\boldsymbol{d}^2$.

Now for the mean reversion parameters. Again, there are multiple approaches. As with Model $\mathcal{B}$, autoregression coefficients can be used. To account for the regime changing dynamics, it makes sense to regress on the past three days. The regression coefficients will be denoted $\xi_0(\cdot), \xi_1(\cdot), \xi_2(\cdot)$ and $\xi_3(\cdot)$. Additionally, autocorrelation coefficients up to lag 3 can also be explored, these will be denoted $\nu_0(\cdot), \nu_1(\cdot), \nu_2(\cdot)$ and $\nu_3(\cdot)$

Finally, the spike parameters μ_1 and p need to be considered. These parameters introduce skew to the trajectory distribution. Statistics that measure skew include $\text{sk}(\cdot)$, the Maximum Likelihood Estimators for a Gamma distribution and the difference between the mean and median. The Maximum Likelihood Estimators for the Gamma distribution will be denoted $\theta_1(\cdot)$ and $\theta_2(\cdot)$ for the shape and rate parameters respectively. The

median will be denoted $m(\cdot)$. Notably, since a Gamma distribution is defined for support $(0, \infty)$, the Maximum Likelihood Estimators will be evaluated for $(\exp(y_i), y_i \in \boldsymbol{y})$, the exponentiated trajectory.

Further, the spike parameters have similar effects to the volatility parameters, thus the statistics mentioned to capture those parameters are also of interest. However, even with all these statistics, it will still be difficult to get a good estimate of p, the probability of remaining in regime 0. A statistic that is directly correlated with the frequency of the spikes is needed. To do this, the statistic denoted $\hat{p}(\cdot)$ will be constructed.

To construct $\hat{p}(\cdot)$, first notice the structure of the spikes. If there is a spike on day i, then the additional stochastic increment should increase the value $|d^1_{i-1}| = |y_i - y_{i-1}|$. However, volatility σ_0 could also cause this. The crucial element is on day $i+1$, the model will be in the reverting regime, so $|d^1_i| = |y_{i+1} - y_i|$ should also be large. This is shown in Figure 12. Thus, construct the spike series $\boldsymbol{s}$ by $s_i = |d^1_{i-1}| + |d^1_i|$. To get an idea for the frequency of spikes, define $\hat{p}(\cdot)$ to be the number of elements s_i of $\boldsymbol{s}$, such that $s_i > \bar{x}(\boldsymbol{s}) + 2\,\mathrm{sd}(\boldsymbol{s})$. This statistic should provide high correlation with the frequency of the spikes, thus, demonstrate high correlation with p.

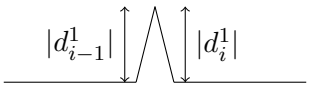

Figure 12: Spike Statistic

These statistics can be investigated using the correlation diagrams from Section 3.5. The results can be seen in Figure 13.

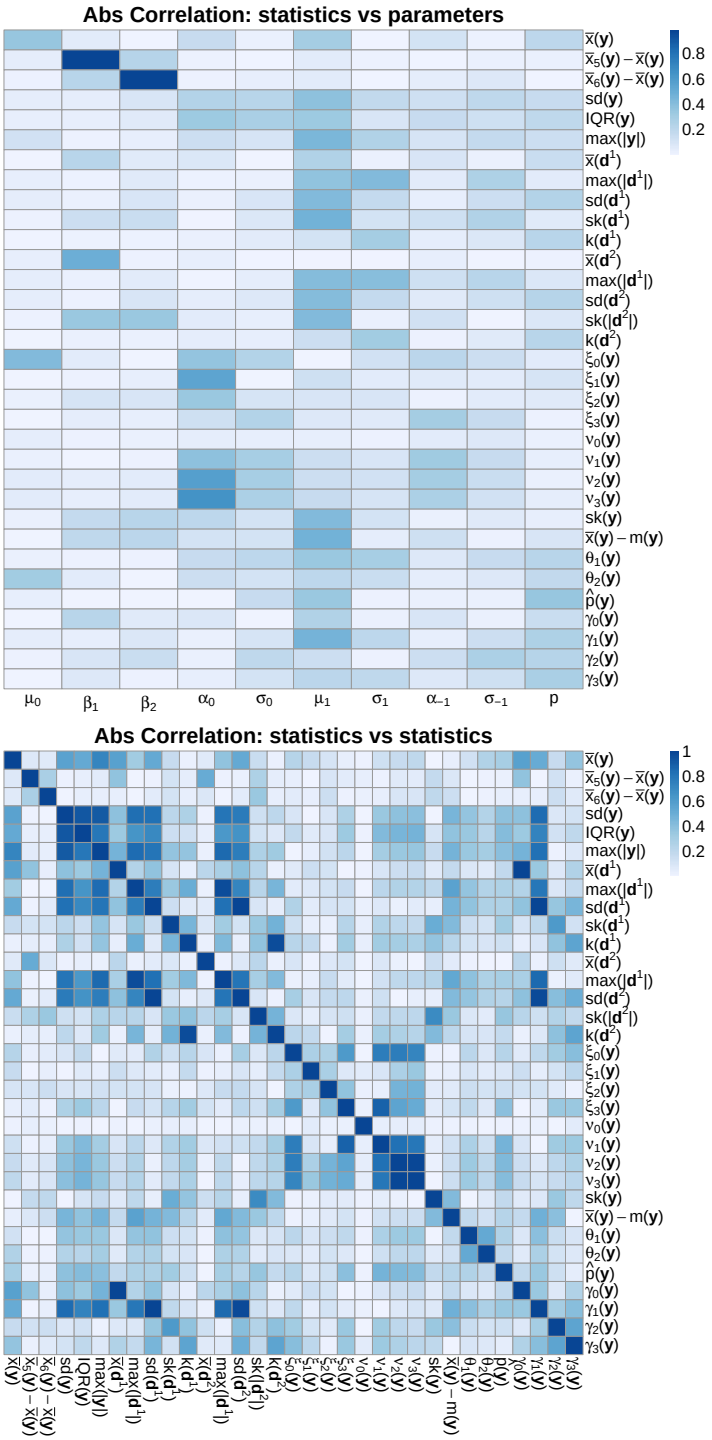

Figure 13: Model $\mathcal{A}$ Correlation Diagrams

Most statistics can be removed due to either having not enough correlation with model parameters or too much correlation with other statistics. In particular, autoregression coefficients are favoured over autocorrelation coefficients due to their smaller mutual correlation, and all statistics based on the second order differences $\boldsymbol{d}^2$ did not demonstrate enough correlation with model parameters to be included.

After some trial and improvement, $\bar{x}$, $\bar{x}_5 - \bar{x}$, $\bar{x}_6 - \bar{x}$, ξ_1, ξ_3, IQR, $\max|\boldsymbol{d}^1|$, $\max|\boldsymbol{y}|$, $\hat{p}$, γ_0, γ_1, γ_2 and γ_3 were found, jointly, to be the best performing statistics.

Synthetic Likelihood Estimation can now be performed for Model $\mathcal{A}$. The next section will check for convergence to the true parameters.

4.2.2 Checking Convergence

Before proceeding with this section, one final adjustment needs to be made. Spike parameter μ_1 needs to be replaced with the transformed parameter τ_{μ_1} where $\mu_1 = \exp\tau_{\mu_1}$. This constrains μ_1 to positive values. This is a useful constraint since the price spikes of interest are positive and negative spikes can upset the Synthetic Likelihood Estimation.

Now for the simulation study. Using the same process used for Model $\mathcal{B}$, a simulation study can be used to check for convergence to the true parameters (Table 1). As before, 50 simulations were conducted. Box plots of the results are found in Figure 14; RMSE values are found in Table 3.

Table 3: Root Mean Squared Errors for Model $\mathcal{A}$

Parameter:	μ_0	β_1	β_2	α_0	σ_0	μ_1	σ_1	α_{-1}	σ_{-1}	p
RMSE (3 s.f.):	0.177	0.0631	0.0560	0.941	3.75	23.7	4.14	0.974	6.29	0.879

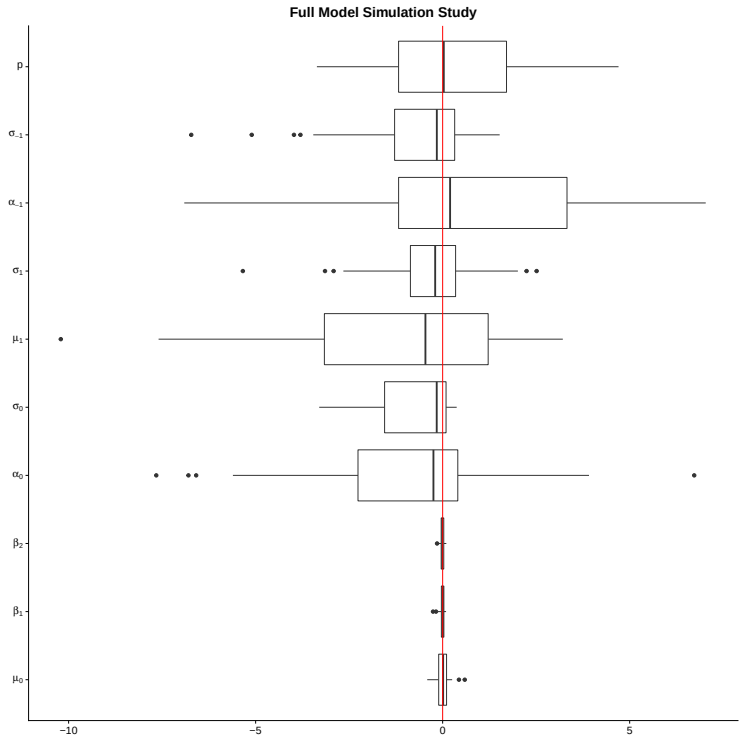

Figure 14: Model $\mathcal{A}$ Simulation Study Box Plots

All parameters converge around their true values. Since the simulation study only used an observation length of 100, tighter fits with fewer outliers would be expected if the observation length was increased (Theorem 3.9). The standout value for the RMSEs, is the value for τ_{μ_1}. This is partly since for an observation length of 100 and a p of 0.95, there would only be approximately five spikes per simulation. This makes approximating the spike parameters a tall order that would be solved by using a longer observation length. This was not computationally feasible for this project.

Since the simulation study can provide strong evidence that the estimation will converge to the true parameters, the next section will cover fitting some real spot price data from Nordpool.

4.2.3 Fitting With Nordpool Data

The data used for this project are the N2EX Day Ahead Auction Prices in GBP available from Nordpool's website.

Since Model $\mathcal{A}$ does not take seasonality into account, this project will look at spot prices for a particular season. Specifically, the log prices from Winter 2020 (October 2020 to March 2021) shown in Figure 15. There are two large spikes in the observed trajectory and maybe three or four smaller spikes. Once the model is fit, these spikes should be observed in the sampled trajectories.

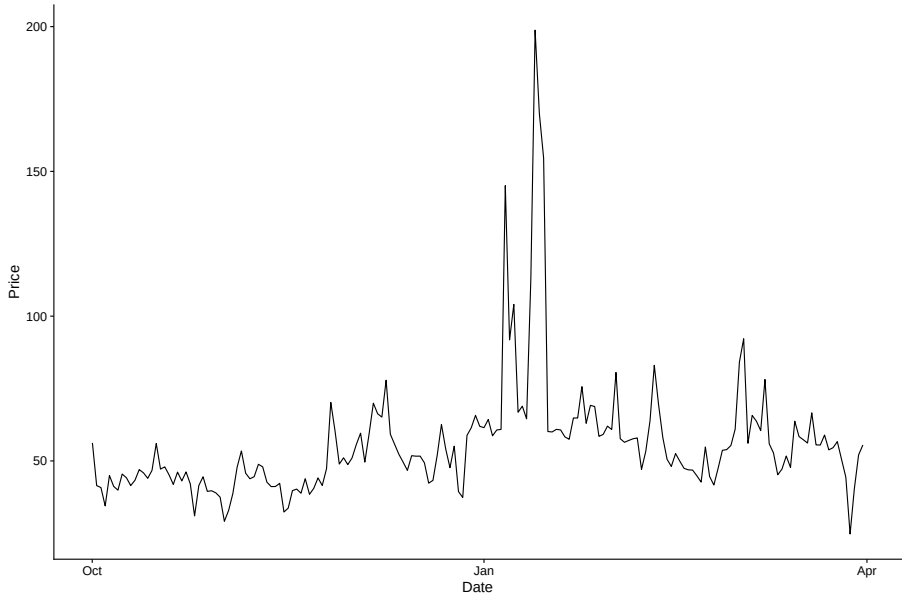

Figure 15: Observed Trajectory

In Table 4, the initial and final parameter values are shown. The initial values were chosen by heuristic and from experience. For μ_0, β_1 and β_2, samples means were used. For α_0, a small amount of mean-reversion was included (10%), for α_{-1} this was increased to 20% to reflect that more mean-reversion would be expected after a spike compared to during regime 0 (default price levels). For σ_0, a very small amount of volatility was included (0.05). This was increased to 0.5 for during and after a spike (σ_1 and σ_{-1} respectively). Finally, for the spike parameters, μ_1 was set to 0.5 to denote a small positive, but noticeable spike and p was set to 0.995 so that initially, the trajectory would remain, for the most part, in regime 0.

Table 4: Fitted Values for Nordpool Data

Parameter	Initial Value (3 s.f.)	Final Value (3 s.f.)
μ_0	$\bar{x}(\boldsymbol{y}^0) = 3.97$	3.93
β_1	$\bar{x}_5(\boldsymbol{y}^0) - \bar{x}(\boldsymbol{y}^0) = 0.0213$	0.0434
β_2	$\bar{x}_6(\boldsymbol{y}^0) - \bar{x}(\boldsymbol{y}^0) = 0.095$	0.109
τ_{α_0}	$\log \frac{0.1}{1-0.1} = -2.20$	-1.01
τ_{σ_0}	$\log 0.05 = -3.00$	-2.13
τ_{μ_1}	$\log 0.5 = -0.693$	-1.68
τ_{σ_1}	$\log 0.5 = -0.693$	-2.15
$\tau_{\alpha_{-1}}$	$\log \frac{0.2}{1-0.2} = -1.39$	1.66
$\tau_{\sigma_{-1}}$	$\log 0.5 = -0.693$	-0.474
τ_p	$\log \frac{0.995}{1-0.995} = 5.29$	3.40

A τ_p of 3.40 corresponds to a p of 0.97. Since the trajectory length is 182, around five spikes per sampled trajectory would be expected. This lines up nicely with the observed trajectory. However, the real test is if the sampled trajectories resemble the observed trajectory. Five sampled trajectories can be seen in Figure 16.

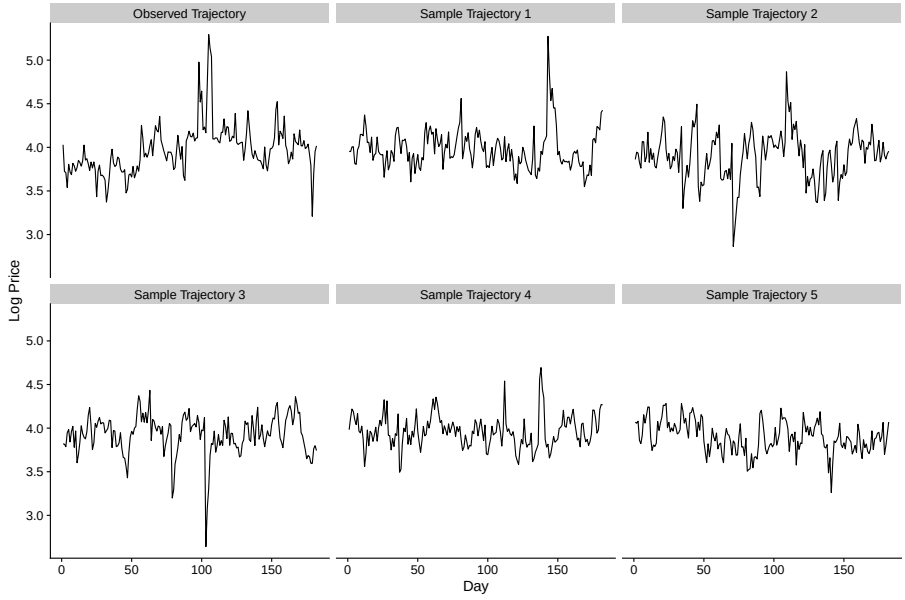

Figure 16: Simulated Trajectories

The trajectories match quite nicely in all but one aspect. The sampled trajectories have some negative spikes. This, as discussed in Section 3, is a result of drawing the spikes from a Normal distribution. It should be noted that since the log price is being considered, these spikes will be less pronounced in the spot prices once the trajectories are exponentiated. Even so, this is an undesirable feature and modifying the model to eliminate these negative spikes presents an opportunity for future projects to explore.

5 Conclusion

This project set out to

1. use 'Synthetic Likelihoods' (Wood 2010) to perform statistical inference for electricity spot prices;
2. evaluate the use of the Robust Covariance Matrix (Huber 1967) as an estimate of the sample covariance calculated during the evaluation of a Synthetic Likelihood.

On the first point, this project can conclude that Synthetic Likelihoods can be used to fit electricity price data. This is backed up from the simulation study results and the trajectories sampled using the Nordpool data. Further, the methods provided can be used under desirable modifications to Model $\mathcal{A}$, such as replacing the Normally distributed spikes with Log-Normal spikes. This modification would not be possible under the Kalman Filter methadology employed in Huisman & Mahieu (2003).

Further, this project has found Synthetic Likelihoods to be a powerful method of statistical inference. The ability to use a collection of heuristics for parameter inference is particularly beneficial. The difficulty is choosing the statistics. This project demonstrated the use of correlation diagrams to quickly obtain an estimate of the quality of a set of statistics. The statistic $\hat{p}$ was proposed, which successfully captured the frequency of the spikes generated by Model $\mathcal{A}$.

On the second point, this project found the Robust Covariance Matrix to provide significant computational benefit. In particular, a $37\times$ reduction in the time to fit Model $\mathcal{B}$ when computing the Hessian analytically. Thus, computing the Hessian by hand is a tedious but computationally desirable step. Additionally, the restriction to M-Estimators is significant and this project was not able to fit Model $\mathcal{A}$ using only M-Estimators.

Whilst both initial aims of the project were fulfilled, there is scope for further work. First, in finding a collection of M-Estimators that can fit Model $\mathcal{A}$ and second in ex-

tending Model $\mathcal{A}$ to include Log-Normal price spikes. Additionally, more work fitting the Nordpool data would also be beneficial. Summer/Winter seasonality and fundamental changes caused by, for example, the rising gas prices in 2021/2022 present challenges in this area that future projects could look into.

As a final remark, both aims of the project were satisfied: Synthetic Likelihoods were found to be effective at fitting models for electricity spot prices and the Robust Covariance Matrix provides significant computational benefit to Synthetic Likelihood Estimation. Further, this project provides fertile ground for future projects to explore, particularly in extending Model $\mathcal{A}$ and fitting the Nordpool data.

References

Fasiolo, M., Pya, N. & Wood, S. N. (2016), 'A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology', *Statistical Science* **31**(1), 96–118.

Geyer, C. J. (2013), '5601 notes: The sandwich estimator', `https://www.stat.umn.edu/geyer/5601/notes/sand.pdf`. [Online; accessed 25-April-2022].

Haben, S., Ward, J., Vukadinovic Greetham, D., Singleton, C. & Grindrod, P. (2014), 'A new error measure for forecasts of household-level, high resolution electrical energy consumption', *International Journal of Forecasting* **30**(2), 246–256.

Hayashi, F. (2000), *Econometrics*, 1 edn, Princeton University Press.

Huber, P. (1967), 'The behavior of maximum likelihood estimates under nonstandard conditions', *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability* **1**, 221–233.

Huisman, R. & Mahieu, R. (2003), 'Regime jumps in electricity prices', *Energy Economics* **25**(5), 425–434.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problems', *Journal of Basic Engineering* **82**(1), 35–45.

Keil, C. & Craig, G. C. (2009), 'A displacement and amplitude score employing an optical flow technique', *Weather and Forecasting* **24**(5), 1297–1308.

Krzanowski, W. J. (2000), *Principles of multivariate analysis*, Oxford University Press.

Matteo Fasiolo and Simon N. Wood (2014), 'synlik-dev', `https://github.com/mfasiolo/synlik_dev`. [Online; Version 0.2.0].

May, R. M. (1976), 'Simple mathematical models with very complicated dynamics', *Nature* **261**(5560), 459–467.

Pilipović, D. (2007), *Energy risk*, 2 edn, McGraw-Hill.

Pollard, D. (1985), 'New ways to prove central limit theorems', *Econometric Theory* **1**(3), 295–313.

Rossi, R. J. (2018), *Mathematical Statistics*, 1 edn, John Wiley & Sons, Inc.

Wood, S. N. (2010), 'Statistical inference for noisy nonlinear ecological dynamic systems', *Nature* **466**(7310), 1102–1104.