Contents

1 Introduction

2 Methodology

2.1 CoDA

2.1.1 Motivation

One way to see our data is as a compositional time series. The exact definition of compositional will follow later but in general compositional data, which is by nature multivariate, describes relations between the parts instead of absolute values. We transform the data in such a way, that the values of each category can be seen as the relative share of the total amount at the current time. We then predict the relative share of the category for the next point in time. Since we are ultimately interested in the absolute value, we include the total sum of all categories as an additional variable and then use it for calculating the absolute value. This is modelled as the so-called $\mathcal{T}$-Space which will be introduced later. Since VAR models are easy to estimate and interpret and have some beneficial properties with our choice of transformation, we opt to focus on them. One such property is the fact that the VAR model does not depend on the concrete choice of the ilr-transformation [KFH15].

2.1.2 Preliminaries

The basis of this section is given by [KFH15], [Ego+03] and [FH20].

CoDA, which is short for "'Compositional Data Analysis"' works with compositional data. The key to compositional data is the fact that the absolute value of its parts is less important than the relative relation of the parts to each other. To define compositional data, we first need to define the $(D-1)$-dimensional simplex,

$$\mathbb{S}^D := \left\{ (x_1, \ldots, x_D)^T : x_i > 0, i = 1, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}$$

where κ is a positive constant [KFH15]. The choice of κ is not relevant, as the relative information in the compositional parts stays the same. A D-dimensional vector $\mathbf{x} = (x_1, \ldots, x_D)^T$ is said to be compositional if it is part of $\mathbb{S}^D$. Next we can induce a $(D-1)$-dimensional vector space on $\mathbb{S}^D$ by perturbation and power transformation. For compositions $\mathbf{x}, \mathbf{y} \in \mathbb{S}^D$ and $\alpha \in \mathbb{R}$ they are defined respectively as [KFH15]

$$\mathbf{x} \oplus \mathbf{y} := \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D)^T, \quad \alpha \odot \mathbf{x} := \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)^T.$$

Here $\mathcal{C}$ is the closure operation that maps each compositional vector from the real value space $\mathbb{R}_+^D$ into its representation in $\mathbb{S}^D$

$$\mathcal{C}\left(\boldsymbol{x}\right) := \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i}\right)^{T}.$$

Using $y^{-1} := \mathcal{C}(y_1^{-1}, y_2^{-1}, \ldots, y_D^{-1})$, the inverse perturbation can be defined as

$$\mathbf{x} \ominus \mathbf{y} := \mathbf{x} \oplus \mathbf{y}^{-1}$$

Now we further define an inner product in order to have an inner product space over the simplex $\mathbb{S}^D$. For two compositions $\mathbf{x}, \mathbf{y} \in \mathbb{S}^D$ define the Aitchison inner product as

$$\langle \mathbf{x}, \mathbf{y} \rangle_a := \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log(\frac{x_i}{x_j}) \log(\frac{y_i}{y_j}).$$

In addition, a norm and distance measure can be defined

$$\|\mathbf{x}\|_a^2 := \langle \mathbf{x}, \mathbf{x} \rangle_a, \quad d_a(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} \ominus \mathbf{y}\|_a.$$

This induced geometry is called the Aitchison geometry and it allows us to express a composition $\mathbf{x} \in \mathbb{S}^D$ as a perturbation-linear combination of a basis of $\mathbb{S}^D$.

However, in order to use standard statistical tools, it is desirable to move from this geometry to the Euclidean real space [FH20]. There are various ways to map the data from the simplex $\mathbb{S}^D$ to the real space $\mathbb{R}^D$. A review of the most common transformations is provided in the following section.

2.1.3 Common Transformations

Let $\mathbf{x}, \mathbf{y} \in \mathbb{S}^D$ be D-part compositions.

alr Coordinates

The additive log-ratio (alr) Coordinates are defined as

$$\mathbf{y}^{(k)} = alr_k(\mathbf{x}) := \left(\log\left(\frac{x_1}{x_k}\right), \ldots, \log\left(\frac{x_{k-1}}{x_k}\right), \log\left(\frac{x_{k+1}}{x_k}\right), \ldots, \log\left(\frac{x_D}{x_k}\right)\right).$$

and map the composition $\mathbf{x}$ to the real space $\mathbb{R}^D$. They are mainly mentioned for historic purposes since they are an intuitive way of transformation. However, limitations are posed by their dependence on the choice of the denominator x_k and the fact that they are not orthogonal to each other [FH20].

clr Coefficients

Let $g(\mathbf{x})$ be the geometric mean of $\mathbf{x}$. The centered log-ratio coefficients are then defined as

$$\mathbf{z} = (z_1, \ldots, z_D)^T = clr(\mathbf{x}) := \left(\log \left(\frac{x_1}{g(\mathbf{x})} \right), \ldots, \log \left(\frac{x_D}{g(\mathbf{x})} \right) \right)^T.$$

This transformation maps $\mathbf{x}$ into the hyperplane $V = \left\{ \mathbf{z} \in \mathbb{R}^D : \sum_{i=1}^{D} z_i = 0 \right\} \subset \mathbb{R}^D$. Hence the transformed data is constrained, which is emphasised by the term 'coefficient' instead of 'coordinates' [FH20]. It can be shown that the clr transformation is an isometry[Ego+03]. Therefore it holds

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle clr(\mathbf{x}), clr(\mathbf{y}) \rangle_a,$$
$$d(\mathbf{x}, \mathbf{y})_a = d(clr(\mathbf{x}), clr(\mathbf{y})).$$

ilr Coordinates

The isometric log-ratio (ilr) are closely related to the clr Coefficients. Assume the inverse clr transformation is isometric. Let $\{v_1, \ldots, v_n\}$ be an orthonormal base in in the D-dimensional hyperplane V. Then $\mathbf{e}_i = clr^{-1}(v_i), i = 1, \ldots, D-1$ is an orthonormal basis of the simplex $\mathbb{S}$. For $\mathbf{x} \in \mathbb{S}$, the ilr transformation can then be defined as

$$\mathbf{u} = ilr(\mathbf{x}) = \left(\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \right)^T.$$

In addition to being isometric, the ilr transformation is also isomorph. Let $\mathbf{x}, \mathbf{y}$ be two compositions and $\alpha, \beta \in \mathbb{R}$. Then,

$$ilr(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha \cdot ilr(\mathbf{x}) + \beta \cdot ilr(\mathbf{y})$$

as well as,

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle ilr(\mathbf{x}), ilr(\mathbf{y}) \rangle_a,$$
$$d(\mathbf{x}, \mathbf{y})_a = d(ilr(\mathbf{x}), ilr(\mathbf{y})),$$
$$\|x\|_a = \|ilr(x)\| = \|u\|.$$

From the definition of the ilr coordinates it can be seen that they can be expressed as a linear combination of the basis induced by the clr coefficients as seen above. Let $\mathbf{V}$ be a $D \times (D-1)$ matrix with columns $\mathbf{v}_i = clr(\mathbf{e}_i)$. For a composition $\mathbf{x}$ the vector of ilr coordinates associated with $\mathbf{V}$ is given by,

$$\mathbf{u_V} = ilr_\mathbf{V}(\mathbf{x}) = \mathbf{V}^T clr(\mathbf{x}) = \mathbf{V}^T \log(\mathbf{x}).$$

The matrix $\mathbf{V}$ is the contrast matrix with the orthonormal basis $(\mathbf{e}_i)_{i=1}^{D-1}$ [Ego+03]. A special choice of orthogonal coordinates leads to the coordinates

$$ilr(\mathbf{x}) = (z_1, \ldots, z_{D-1})^T,$$

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \log\left(\frac{x_j}{\sqrt[D-j]{\prod_{l=j+1}^{D} x_l}}\right), \quad j = 1, \ldots, D-1.$$

With this choice, the problem of interpretation, which arises from the relative nature of the compositional data and the dimension of the simplex, can be solved. The part x_1 is only contained in z_1 and therefore contains all relative information of x_1 [FH20].

To transform the data back in the simplex, the inverse transformation is given by,

$$x_1 = \exp\left(\sqrt{\frac{D-1}{D}} z_1\right),$$

$$x_i = \exp\left(\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \sqrt{\frac{D-i}{D-i+1}} z_i\right), \quad i = 2, \ldots, D-1,$$

$$x_D = \exp\left(-\sum_{j=1}^{D-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j\right).$$

2.1.4 The VAR Model

Since we have established the basic setting we can now introduce compositional time series (CTS). A CTS $\{\boldsymbol{x}_t : t = 1, \ldots, n\}$ can be defined as a series where $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{Dt})^T \in \mathbb{S}^D$. They are thus characterised by their positive components which sum up to a constant κ for each point in time $t = 1, \ldots, n$

$$\sum_{i=1}^{D} x_{it} = \kappa, \quad x_i > 0, i = 1, \ldots D \, ; t = 1, \ldots, n.$$

Let $\mathbf{x}_t$ be a D-dimensional compositional vector measured at time t, $t = 1, \ldots, n$ and $\mathbf{z}_t = ilr(x_t)$ be its ilr transformation determined by the matrix $\mathbf{V}$. Then the VAR model with lag order p is given by [KFH15]

$$\mathbf{z}_t = \mathbf{c_V} + \mathbf{A_V}^{(1)} \mathbf{z}_{t-1} + \mathbf{A_V}^{(2)} \mathbf{z}_{t-2} + \ldots + \mathbf{A_V}^{(p)} \mathbf{z}_{t-p} + \boldsymbol{\epsilon}_t. \tag{2.1}$$

where $\mathbf{c_V} \in \mathbb{R}^{D-1}$ is a real vector, $\mathbf{A_V}^{(i)} \in \mathbb{R}^{(D-1)\times(D-1)}$ are parameter matrices and $\boldsymbol{\epsilon}_t$ is a white noise process with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The observation $\mathbf{z}_t$ therefore depends on the p past observations $\mathbf{z}_{t-1}, \ldots, \mathbf{z}_{t-p}$. It can be shown, that two VAR(p) models resulting from different ilr transformations are compositionally equivalent which means that the same predictions are obtained [KFH15].

Estimation of the VAR Model

Assuming a sample size of n, equation 2.1 can be written in matrix form as

$$\mathbf{Y} = \mathbf{Z}\mathbf{B} + \mathbf{E},$$
$$\mathbf{Y} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T \in \mathbb{R}^{n \times (D-1)},$$
$$\mathbf{Z} \in \mathbb{R}^{n \times [(D-1)p+1]} \text{ with } \mathbf{Z}_t = \left(1, \mathbf{z}_{t-1}^T, \ldots, \mathbf{z}_{t-p}^T\right)^T,$$
$$\mathbf{B} = \left[\mathbf{c}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(p)}\right]^T \in \mathbb{R}^{(D-1)p+1 \times (D-1)}.$$

The parameter $\mathbf{B}$ can then be estimated separately for each column of $\mathbf{Y}$ by the ordinary least squares (OLS) method. In addition, if there are no restrictions posed on the parameter, the estimator is equal to the generalised least squares (GLS). If the VAR(p) process is normally distributed and the rows of the error matrix $\mathbf{E}$ represent a white noise process, thus $\boldsymbol{E} \sim WN(\Sigma)$ where Σ is the covariance matrix, then the estimator is also equal to the maximum likelihood (ML) estimator. Under these assumptions it can be shown that the OLS estimator is consistent and asymptotic normal [KFH15] [Lüt07].

2.1.5 $\mathcal{T}$-Spaces

Since in our context absolute values should be predicted eventually, we model this information in form of the sum of the original parts. Let $\mathbf{x}$ be a D-dimensional compositional vector and define an extended vector space $\mathcal{T} = \mathbb{R}_+ \times \mathbb{S}^D$. An element of $\mathcal{T}$ now has the form $\tilde{\boldsymbol{x}} = [t(\boldsymbol{x}), \mathcal{C}(\boldsymbol{x})] = [t_x, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D]$ where $t(\boldsymbol{x}) = \sum_{i=1}^{D} x_i$. This allows us to model the relative structure of the data as well as the total sum of the compositional parts. Often times the logarithm of the sum is taken. In the subsequent analysis, the compositional part $\boldsymbol{x}$ is then transformed using one of the possible transformations and then the statistical analysis is performed. In our case, back transformation of the results are required as well as scaling the proportional values back to the absolute values to get a final result. [KFH15]

2.2 INGARCH

In this section we introduce the INGARCH(p,q) model. First we provide a motivation on why we chose this model and review some other possible models for discrete time series count data. The review is mainly based on [Lib16] and [Hei03]. A more detailed review can be found in [MMM97]. Subsequently we define the INGARCH(p,q) model itself and list some of its properties.

2.2.1 Motivation

This section is based on [Lib16] and [Hei03].
Since our data can be seen as a discrete time series with count data, we want a model

which is able to take these properties into account. In addition, autocorrelation and overdispersion are two common features in count data. One common way to deal with count data are Markov chains. The dependent variable can take on all possible values in the so called state space and the probability of changing states is then modelled as a transition probability. A limitation is the fact that these models become cumbersome if the state space gets too big and lose tractability. As an extension to the basic Markov chains models, Hidden Markov chains are proposed by [MMM97]. However, since there is no generally accepted way to determine the order of this model, it can cause problems if the data structure does not provide intuitive ways to do it. Another issue is that the number of parameters which need to be estimated gets big quickly, especially if the order of the model is big.

Other common models for time series data are the ARMA models. There exists a discrete version of them in the form of the Discrete Autoregressive Moving Average (DARMA) models. They can be defined as a mixture of discrete probability distributions and a suitable chosen marginal probability function [BS09]. While there have been various applications, for example in [CDK87], there seem to be difficulties in their estimation [Hei03].

State space models with conjugated priors are proposed by [HF89]. The observations are assumed to be drawn from a Poisson distribution whose mean itself follows a Gamma distribution. The parameters of the Gamma distribution are chosen in such a way that its mean is constant but its variance is increasing. While there are ways proposed by [Zeg88] to handle overdispersion, these models have the weakness of needing further assumptions to handle zeros while also having more complicated model specifications [Hei03].

While there are many more possible models, we decided to focus on the class of Generalised Linear Models (GLM). In the case of discrete time series with count data, the observations are modelled conditionally on the past and follow a discrete distribution. The conditional mean is then connected with a link function to the past observations and conditional means. Furthermore, a covariate vector can be introduced to account for external influence. While being easy to use and estimate they still provide a good amount of flexibility. In addition, a wide array of tools is available for various tests and forecasts. From the class of the GLMs we compare the INGARCH(p,q) and a log-linear model, which will be discussed in section 2.3. We then chose the INGARCH(p,q) model based on its superior performance and stability.

2.2.2 INGARCH Model

Denote a time series consisting of integer values with $\{Y_t : t \in \mathbb{N}\}$ and a r-dimensional time varying covariate vector with $\mathbf{X}_t = (X_{t,1}, \ldots, X_{t,r})^T$. Let the conditional mean be $\lambda_t = \mathbb{E}[Y_t|\mathcal{F}_{t-1}]$ where $\mathcal{F}_{t-1}$ is the sigma-field generated by Y_l and λ_l for $l < t$, $\mathcal{F}_{t-1} = \sigma(Y_1, \ldots, Y_l, \lambda_1, \ldots, \lambda_l)$. Therefore, the conditional mean of the time series is dependent on its combined history of the past conditional means and its past values. With this, we can define the integer valued generalized autoregressive conditional heteroskedasticity model of order (p,q) (INGARCH(p,q) model) as,

$$Y_t|\mathcal{F}_{t-1} \sim P(\lambda_t); \forall t \in \mathbb{N}, \tag{2.2}$$

$$\mathbb{E}\left[Y_t|\mathcal{F}_{t-1}\right] = \lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k} + \sum_{l=1}^{q} \alpha_l \lambda_{t-l} \tag{2.3}$$

where $p, q \in \mathbb{N}$ and $P(\lambda_t)$ is a Poisson distribution with mean λ_t. The integer p defines the number of past values to regress on, whereas q does the same for the past conditional means. In order to account for external effects as well, we add the covariate vector $\mathbf{X}_t$

$$Y_t|\mathcal{F}_{t-1} \sim P(\lambda_t); \forall t \in \mathbb{N}, \tag{2.4}$$

$$\mathbb{E}\left[Y_t|\mathcal{F}_{t-1}\right] = \lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k} + \sum_{l=1}^{q} \alpha_l \lambda_{t-l} + \boldsymbol{\eta}^T \mathbf{X}_t \tag{2.5}$$

where $\boldsymbol{\eta}$ is the parameter for the covariates.

The distributional assumptions $Y_t|\mathcal{F}_{t-1} \sim P(\lambda_t)$ implies

$$p_t(y; \boldsymbol{\theta}) = \mathbb{P}(Y_t = y|\mathcal{F}_{t-1}) = \frac{\lambda_t^y \exp(-\lambda_t)}{y!}, \quad y \in \mathbb{N}_0. \tag{2.6}$$

Furthermore it can be shown that conditionally on the past history $\mathcal{F}_{t-1}$ the model is equidispersed, i.e. it holds $\lambda_t = \mathbb{E}\left[Y_t|\mathcal{F}_{t-1}\right] = \mathbb{V}\left[Y_t|\mathcal{F}_{t-1}\right]$. However, unconditionally the model exhibits overdispersion. In that case it holds $\mathbb{E}\left[Y_t\right] \leq \mathbb{V}\left[Y_t\right]$ [Hei03].

Estimation of the INGARCH Model

We summarise the estimation of the INGARCH(p,q) Model as described in [Lib16].

The parameter space for the INGARCH(p,q) model with external effects 2.4 is given by

$$\Theta = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{p+q+r+1} : \beta_0 > 0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1 \right\}.$$

To ensure positivity of the conditional mean λ_t, the intercept β_0 must be positive while all other parameters must be non negative. The upper bound of the sum ensures that the model has a stationary and ergodic solution with moments of any order [FLO06; FRT09; DFT12]. A quasi maximum likelihood approach is used to estimate the parameters $\boldsymbol{\theta}$. For observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ the conditional quasi log-likelihood function, up to a constant, is given by,

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log p_t(y_t; \boldsymbol{\theta}) = \sum_{t=1}^{n} \left(y_t \log(\lambda_t(\boldsymbol{\theta})) - \lambda_t \theta \right). \tag{2.7}$$

where $p_t(y_t; \boldsymbol{\theta})$ is the probability density function defined in 2.6. The conditional mean is seen as a function $\lambda_t : \Theta \to \mathbb{R}^+$. The conditional score function is given by,

$$S_n(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{2.8}$$

The vector $\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ can be computed recursively. The conditional information matrix is given by,

$$\begin{aligned} G_n(\theta) &= \sum_{t=1}^{n} Cov \left(\frac{\partial \ell(\boldsymbol{\theta}; Y_t)}{\partial \boldsymbol{\theta}} \bigg| \mathcal{F}_{t-1} \right) \\ &= \sum_{t=1}^{n} \left(\frac{1}{\lambda_t(\boldsymbol{\theta})} \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T. \end{aligned}$$

Finally, assuming that the quasi maximum likelihood estimator (QMLE) $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ exists, it is the solution to

$$\hat{\boldsymbol{\theta}} := \hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} (\ell(\boldsymbol{\theta})). \tag{2.9}$$

2.3 Other Methods

2.3.1 Naive Random Walk

This method acts as kind of benchmark model and is what is currently used for forecasting. Let $\{Y_t : t \in \mathbb{N}\}$ be again our timeseries. Then

$$\hat{Y}_{t+1} = Y_t, \quad \forall t \in \mathbb{N} \tag{2.10}$$

where $\hat{Y}_t$ is the predicted value at time t. In other words, the last known value is the predicted value.

3 Application

3.1 Model Specifications

As our data has a specific structure, some transformations can be made to increase performance and stability. The most prominent characteristic of our data is its amount of 0 or null values. As both CoDA and INGARCH can't handle an excessive amount of 0 values, we have to accommodate for this. The concrete way to do this will be described in the following subsections.

Another varying factor is the length of the history. We define as history the length of the timeseries. While at first it may seem obvious to use as much data as possible, it may actually not always result in a better model. Older values may contain outdated information which influences the estimation the of parameters. Therefore we compare the performance of the models with various history lengths.

Closely related to the length of the history, is the shape of the window used. The window determines which values are used to estimate the parameters at each point in time. This includes both the initial length of it and the way it handles new values. As the different timeseries vary in length, we choose the possible window length as a fraction of the timeseries history. For the way how new values are handled, we focused on two different approaches. The first one uses a fixed window length. This means when a new time point is available, it will be included in the estimation while simultaneously the oldest time point will be removed from the estimation. This has the advantage of only using the most recent and relevant information. The second approach, extends the window at each point in time. When a new value is available, it is included in the estimation of the parameter. With this approach we have more data available at each step and combined with the varying history length we don't have to rely on information that is too old.

3.1.1 CoDA Specifications

As mentioned above the CoDA model must not include any zero values. Since in the CoDA context we see our data as relative data, a value of zero is not defined. Therefore we need to replace them. In order to keep things simple, we consider two options. The first one adds 0.5 to all timeseries values. The second one only replaces zero values with 0.5.

As already hinted in the description of the methodology we consider the use of $\mathcal{T}$-Spaces. For this, at each time point, we calculate the total amount and include it as an additional variable in the model. In addition we can choose to take the logarithm of the

sum.

Another characteristic of our data are the low values for some categories of it. Even at the aggregated main category level there are instances with low values for some of the categories. This is the case especially for category 3 and 4. As such, we inspected a method which we call in the following one-vs-all. The principle is the following. A category is chosen as the pivot category. For all the chosen time points, at each point, the values of the other categories get summed up. Together with the pivot category, the sum of the other categories are then transformed as usual and the VAR model is calculated. All categories are chosen as a pivot category at one point and the predicted values of the pivot groups are then used as the final result.

3.1.2 INGARCH Specifications

As with the CoDA model, we need to take care of the zero values. Again to keep things simple, we replace missing and zero values with 1.

As an alternative to the Poisson distribution in 2.6, a negative binomial distribution can be used as well. This would change 2.6 to

$$p_t(y; \boldsymbol{\theta}) = \mathbb{P}(Y_t = y | \mathcal{F}_{t-1}) = \frac{\Gamma(\phi + y)}{\Gamma(y + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda_t} \right)^{\phi} \left(\frac{\lambda_t}{\phi + \lambda_t} \right)^{y}, \ \ y \in \mathbb{N}_0.$$

With the negative Binomial Distribution the conditional variance can be larger than the conditional mean $\lambda_t = \mathbb{V}[Y_t | \mathcal{F}_{t-1}] > \mathbb{E}[Y_t | \mathcal{F}_{t-1}]$.

As seen in the model **??** we can also choose too include external factors or not. However, as our data is of the structure where we don't have information about $\boldsymbol{X}_t$ at time t, we cannot make use of it. The values p and q are also varying parameters which have to be chosen.

3.1.3 Error Measure

In order to compare the results of the methods with each other we will introduce a new error measure. The goal of this measure is to get a performance indicator for each fridge which can then be compared and summarised. Since the scales of the fridges vary, the measure should be scale independent. Because our data contains many zeros, we cannot use a percentage error measure. In addition we want to penalise big absolute difference between the predicted values and actual values. This leads us to the following measure.

Let $i = 1, \ldots, T$ denote the point in time and $k = 1, \ldots, K$ the category. Then y_{ik} is the $i - th$ true value of the time series for category k, $\hat{y}_{ik}$ the predicted value and $y_{n_{ij}}$ the naive predicted value . Then we define our measure as

$$\frac{\sum_{k=1}^{K} \sum_{i}^{T} (y_{ik} - \hat{y}_{ik})^2}{\sum_{k=1}^{K} \sum_{i}^{T} (y_{ik} - y_{naive_{ik}})^2}. \tag{3.1}$$

With the use of the squared difference we penalise big deviations from the true value. By taking the naive random walk model as a benchmark, we achieve scale independence

and are able to compare the performance of our model over different timeseries. This error measure is basically the ration of the mean MSEs for the chosen model and the naive random walk model

$$\frac{\frac{1}{K}\sum_{k=1}^{K} MSE_k}{\frac{1}{K}\sum_{k=1}^{K} MSE_{naive_k}}. \tag{3.2}$$

If the ratio is below 1, the mean of the MSEs of our methods is lower than that of the naive method and vice versa. This provides a performance indicator for our models.

Extension of the Error Measure

The measure in 3.1 can be further extended. For example, by allowing to use a subset of all possible categories instead of all. Let $G_K \subset \{1, \ldots, K\}$ then

$$\frac{\sum_{k \in G_K} \sum_i^T (y_{ik} - \hat{y}_{ik})^2}{\sum_{k \in G_K} \sum_i^T (y_{ik} - y_{naive_{ik}})^2}. \tag{3.3}$$

This allows us to compare the performance on the subset of categories over various fridges.

Another possible extension is to take the square root

$$\frac{\sum_{k=1}^{K} \sqrt{\sum_i^T (y_{ik} - \hat{y}_{ik})^2}}{\sum_{k=1}^{K} \sqrt{\sum_i^T (y_{ik} - y_{naive_{ik}})^2}}. \tag{3.4}$$

One future extension which can be investigated is the introduction of weights. This could be used for example when the performance of the model in one category should be put more into focus.

Bibliography

Biswas, Atanu and Peter X.-K. Song. 'Discrete-valued ARMA processes'. In: *Statistics & Probability Letters* 79.17 (2009), pp. 1884–1889. ISSN: 0167-7152. DOI: `https://doi.org/10.1016/j.spl.2009.05.025`. URL: `https://www.sciencedirect.com/science/article/pii/S0167715209001977`.

Chang, Tiao J., J.W. Delleur and M.L. Kavvas. 'Application of Discrete Autoregressive Moving Average models for estimation of daily runoff'. In: *Journal of Hydrology* 91.1 (1987), pp. 119–135. ISSN: 0022-1694. DOI: `https://doi.org/10.1016/0022-1694(87)90132-6`. URL: `https://www.sciencedirect.com/science/article/pii/0022169487901326`.

Doukhan, Paul, Konstantinos Fokianos and Dag Tjøstheim. 'On weak dependence conditions for Poisson autoregressions'. In: *Statistics & Probability Letters* 82.5 (2012), pp. 942–948. ISSN: 0167-7152. DOI: `https://doi.org/10.1016/j.spl.2012.01.015`. URL: `https://www.sciencedirect.com/science/article/pii/S0167715212000259`.

Egozcue, Juan Jose et al. 'Isometric Logratio Transformations for Compositional Data Analysis'. In: *Mathematical Geology* 35 (Apr. 2003), pp. 279–300. DOI: `10.1023/A:1023818214614`.

Ferland, René, Alain Latour and Driss Oraichi. 'Integer-Valued GARCH Process'. In: *Journal of Time Series Analysis* 27.6 (2006), pp. 923–942. DOI: `https://doi.org/10.1111/j.1467-9892.2006.00496.x`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9892.2006.00496.x`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.2006.00496.x`.

Filzmoser, Peter and Karel Hron. '2.30 - Compositional Data Analysis in Chemometrics'. In: *Comprehensive Chemometrics (Second Edition)*. Ed. by Steven Brown, Romà Tauler and Beata Walczak. Second Edition. Oxford: Elsevier, 2020, pp. 641–662. ISBN: 978-0-444-64166-3. DOI: `https://doi.org/10.1016/B978-0-12-409547-2.14591-3`. URL: `https://www.sciencedirect.com/science/article/pii/B9780124095472145913`.

Fokianos, Konstantinos, Anders Rahbek and Dag Tjøstheim. 'Poisson Autoregression'. In: *Journal of the American Statistical Association* 104.488 (2009), pp. 1430–1439. DOI: `10.1198/jasa.2009.tm08270`. eprint: `https://doi.org/10.1198/jasa.2009.tm08270`. URL: `https://doi.org/10.1198/jasa.2009.tm08270`.

Harvey, A. C. and C. Fernandes. 'Time Series Models for Count or Qualitative Observations'. In: *Journal of Business & Economic Statistics* 7.4 (1989), pp. 407–417. ISSN: 07350015. URL: `http://www.jstor.org/stable/1391639` (visited on 04/04/2023).

Heinen, Andreas. *Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model*. MPRA Paper 8113. University Library of Munich, Germany, July 2003. URL: `https://ideas.repec.org/p/pra/mprapa/8113.html`.

Kynčlová, Petra, Peter Filzmoser and Karel Hron. 'Modeling Compositional Time Series with Vector Autoregressive Models'. In: *Journal of Forecasting* 34.4 (2015), pp. 303–314. DOI: `https://doi.org/10.1002/for.2336`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2336`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2336`.

Liboschik, Tobias. 'Modeling count time series following generalized linear models'. In: 2016.

Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, 2007. ISBN: 9783540262398. URL: `https://books.google.at/books?id=muorJ6FHIiEC`.

Macdonald, Lain, Iain L. MacDonald and Iain L. MacDonald. 'Hidden Markov and Other Models for Discrete- valued Time Series'. In: 1997.

Zeger, Scott L. 'A Regression Model for Time Series of Counts'. In: *Biometrika* 75.4 (1988), pp. 621–629. ISSN: 00063444. URL: `http://www.jstor.org/stable/2336303` (visited on 04/04/2023).

List of Figures

List of Tables