Principal component regression in GAMLSS applied to Greek-German government bond yield spreads

Mikis D. Stasinopoulos [1], Robert A. Rigby [1], Nikolaos Georgikopoulos [2] and Fernanda De Bastiani [3]

[1] London Metropolitan University, London, United Kingdom

[2] IOMS, STERN School of Business, NYU, New York, USA

[3] Universidade Federal de Pernambuco, Recife, PE, Brazil

Address for correspondence: Mikis D. Stasinopoulos, School of Computing and Digital Media, London Metropolitan University, 166-220 Holloway Rd, London N7 8DB, United Kingdom.

E-mail: d.stasinopoulos@londonmet.ac.uk.

Phone: (+44) 20 7133 4638.

Fax: (+44) 20 7133 4149.

Abstract: A solution to the problem of having to deal with a large number of interrelated explanatory variables within a generalized additive model for location, scale, and shape (GAMLSS) is given here using as an example the Greek-German government bond yield spreads from the 25th of April 2005 to the 31th of March 2010. Those were turbulent financial years, and in order to capture the spreads

behaviour, a model has to be able to deal with the complex nature of the financial indicators used to predict the spreads. Fitting a model, using principal components regression of both main and first order interaction terms, for all the parameters of the assumed distribution of the response variable seems to produce promising results.

Key words: Box-Cox t; financial spreads; kurtosis; skewness.

1 Introduction

The current paper extends mean and dispersion modelling, where both the location parameter (often the mean) and the scale parameter (often the dispersion) of the distribution of the response variable are modelled as functions of the explanatory variables. Murray Aitkin is one of the pioneers of simultaneously modelling mean and dispersion, Aitkin (1987). His paper "Modelling variance heterogeneity in normal regression using `GLIM`" was one of the few early examples of modelling simultaneously the distribution parameters of a response variable. He proposed the model $y_i = \boldsymbol{\beta}^\top \mathbf{x_i} + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma_i^2)$ where $\text{var}(\epsilon_i) = \sigma_i^2 = \exp(\boldsymbol{\lambda}^\top \mathbf{z}_i)$ for $i = 1, \ldots, n$, and where $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are the coefficients and $\mathbf{x}$ and $\mathbf{z}$, the explanatory variables for modelling the mean and variance of the response variable y_i, respectively. Here we rewrite Murray Aitkin's model as:

$$\begin{aligned} \mathbf{y} &\overset{\text{ind}}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\sigma}) \\ g_1(\boldsymbol{\mu}) &= \mathbf{X}_1 \boldsymbol{\beta}_1 \\ g_2(\boldsymbol{\sigma}) &= \mathbf{X}_2 \boldsymbol{\beta}_2, \end{aligned} \tag{1.1}$$

where the elements of the response variable $\mathbf{y}$ are assumed to be independently (i.e $\overset{\text{ind}}{\sim}$) normally distributed with mean vector $\boldsymbol{\mu}$ and standard deviation vector $\boldsymbol{\sigma}$, and where both the predictors of the mean and standard deviation are linear functions of the explanatory variables, here represented by the design matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. The functions g_1 and g_2 represent known link functions which for the normal distribution are usually set to be the 'identity' and 'log' link functions, respectively. Note that Murray Aitkin used the variance $\boldsymbol{\sigma}^2$ rather the standard deviation $\boldsymbol{\sigma}$ but with a log link the two models are equivalent, since $\log \boldsymbol{\sigma}^2 = 2 \log \boldsymbol{\sigma}$ and therefore any model for $\boldsymbol{\sigma}^2$ is proportional to a model for $\boldsymbol{\sigma}$. In addition, plots for σ are more attractive to human eye than plots for σ^2; see for example Figure 3(c) and (d).

In this paper we consider the following generalization of the Aitkin model:

$$\begin{aligned} \mathbf{y} &\overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\ g_1(\boldsymbol{\mu}) &= \mathbf{T}_{(1,\lambda_1)}\boldsymbol{\gamma}_1 \\ g_2(\boldsymbol{\sigma}) &= \mathbf{T}_{(2,\lambda_2)}\boldsymbol{\gamma}_2 \\ g_3(\boldsymbol{\nu}) &= \mathbf{T}_{(3,\lambda_3)}\boldsymbol{\gamma}_3 \\ g_4(\boldsymbol{\tau}) &= \mathbf{T}_{(4,\lambda_4)}\boldsymbol{\gamma}_4, \end{aligned} \tag{1.2}$$

where now $\mathcal{D}$ represents any theoretical distribution with up four parameters, and where $\boldsymbol{\mu}$ is a vector of location parameters, $\boldsymbol{\sigma}$ is a vector of scale parameters, and $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are vectors of shape parameters of the distribution of the response which often (but not always) model skewness and kurtosis. In this paper, the matrices $\mathbf{T}_{(i,\lambda_i)}$ for $i = 1, 2, 3, 4$ represent the first λ_i principal components of the original design matrices $\mathbf{X}_i$ for $i = 1, 2, 3, 4$. The model given in (1.2) is a special case of generalized additive models for location, scale, and shape (GAMLSS), Rigby and Stasinopoulos (2005), where the numbers of singular vectors λ_i included in each $\mathbf{T}_i$

(a crucial part of modelling the distribution parameters of the response) are the 'tuning' parameters of the model. A general definition of GAMLSS models can be found in Chapter 3 of Stasinopoulos et al. (2017). There are more than 100 available distributions $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ implemented in the package `gamlss.dist` in `R`, that can be found in Rigby et al. (2019). A practical tutorial of using GAMLSS can be found in Stasinopoulos et al. (2018).

This paper is organised as follows. Section 2 describes the motivating example for developing model (1.2). Section 3 shows how the principal component regression model is implemented within GAMLSS. The model building process and the interpretation of the model are shown in Section 4. Conclusions are discussed in Section 5.

2 Greek-German government bond yield spreads

The yields-to-maturity of euro government bonds were and are of great interest and are used as indicators of the financial stability of the Euro zone. With the birth of the European Monetary Union (EMU) many economists and market analysts expected that there would be a permanent reduction in the differences between yields-to-maturity of euro denominated government bonds (with common characteristics, but issued by different EMU countries). Specifically it was expected that each individual EMU country's government bond yields would converge to those of the corresponding German government bond (which was considered the de facto benchmark bond). Unfortunately, and contrary to expectations, during and after the financial crisis of 2007-2008 there was a departure from the (relatively) low yield differences, as these differences exhibited higher levels and acute fluctuations. In this paper we use as our

response variable the *Greek spreads*, that is, the difference between the 10-year Greek government bond yields and the corresponding German bonds. The Greek-German spreads for the period from the 25th of April 2005 to the 31th of March 2010 are shown in Figure 1. There are 2188 observations. This figure shows that at the beginning, the yield difference between the Greek 10-year government bonds and the corresponding German bonds is at a low value and almost at a fixed rate. By the end of 2008 the yield differences start rising, while also the series exhibits acute fluctuations. After May 2010 (with the implementation of the bailout of Greece) the long term Greek government bond market(s) virtually ceased to exist. In this paper we

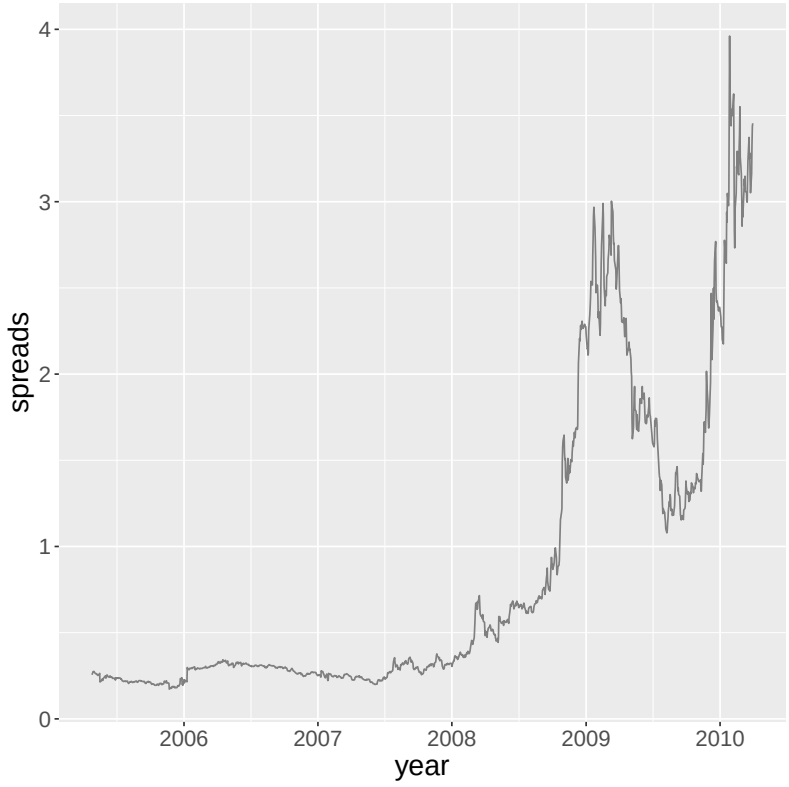

Figure 1: Showing the Greek-German 10 years bond yield spreads during the period from 25th of April 2005 to the 31th of March 2010

will try to model the Greek-German spreads as a function of 67 financial indicators

which we are using as explanatory variables. These variables fall into one or more of following four categories i) *sovereign risk*, ii) *debt level*, iii) *liquidity*, and iv) *volatility*. Those four categories of variables are believed to be important to explain spreads in general.

3 Principal component regression within GAMLSS.

Principal component regression (PCR) has been a statistical tool for a long time. Hotelling (1957) and Kendall (1957) recommended replacing the original explanatory variables in a multiple regression model with their principal components. PCR is one of the techniques examined by Hastie et al. (2009), pp 79, as a supervised statistical learning tool. PCR can be seen as a three-stage procedure. In the first stage, the principal component scores of a (suitably scaled) design matrix are taken. At the second stage a regression is performed treating the principal components scores as the new explanatory variables. At the third stage, to facilitate the interpretation of the model, the fitted coefficients from the PCR can be transferred back to the original design matrix coefficients.

Within a GAMLSS model, let $\mathbf{X}_i$ represent the four different design matrices, of dimension $n \times r_i$ for $i = 1, 2, 3, 4$ for the vectors of parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$, respectively, where n is the number of observations. For simplicity and without loss of generality we shall drop the subscript i and assume that the design matrices for each parameter are identical to $\mathbf{X}$ and of dimensions $n \times r$. Further we will assume that the design matrix $\mathbf{X}$ contains columns of continuous variables which are appropriately *scaled* (in our case, with zero mean and standard deviation equal to one). The

dimensions n, the number of observations, and r, the number of continuous variables, play an important role in what it follows, and often we have to distinguish between the situations when $n > r$ and when $n \leq r$.

3.1 The PCR model

Let $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top$ be the singular value decomposition of the design matrix $\mathbf{X}$, such that, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$ and $\boldsymbol{\Delta}$ be a diagonal matrix containing the singular values for $\mathbf{X}$. When $n > r$ the matrix $\mathbf{U}$ is rectangular with dimensions $n \times r$ while $\boldsymbol{\Delta}$ and $\mathbf{V}$ are squares matrices with dimensions $r \times r$. When $n < r$ the matrices $\mathbf{U}$ and $\boldsymbol{\Delta}$ are square matrices with dimensions $n \times n$ while $\mathbf{V}$ is rectangular with dimensions $r \times n$. The linear space generated by the columns of $\mathbf{X}$ is the same as the linear space generated by the columns of $\mathbf{U}$, i.e. $\mathcal{M}(\mathbf{X}) = \mathcal{M}(\mathbf{U})$, where $\mathcal{M}(\mathbf{A})$ denotes the linear manifold generated by the columns of a matrix $\mathbf{A}$. Also the linear space generated by the rows of $\mathbf{X}$ is the same as the linear space generated by the rows of $\mathbf{V}$ i.e. $\mathcal{M}(\mathbf{X}^\top) = \mathcal{M}(\mathbf{V})$. The principal components (or *scores*) $\mathbf{T}$ of the matrix $\mathbf{X}$ are defined as $\mathbf{T} = \mathbf{X}\mathbf{V} = \mathbf{U}\boldsymbol{\Delta}$, while the matrix $\mathbf{P} = \mathbf{V}^\top$ is called the *loadings*, and $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top = \mathbf{T}\mathbf{P}$.

Let us consider for the moment the case in which $n > r$. Since the matrix of scores $\mathbf{T}$ spans the same linear space as the original matrix $\mathbf{X}$, i.e. $\mathcal{M}(\mathbf{X}) = \mathcal{M}(\mathbf{T})$, any linear (unweighted normal error) regression of the response variable $\mathbf{y}$ into $\mathbf{X}$ or into $\mathbf{T}$ should produce identical fitted values. Let us denote the coefficients for those two regressions as $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, then we have $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ and $\hat{\boldsymbol{\gamma}} = \left(\mathbf{T}^\top\mathbf{T}\right)^{-1}\mathbf{T}^\top\mathbf{y}$ where $\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}$. In addition because the columns of the $\mathbf{T}$ are orthogonal, the estimated $\boldsymbol{\gamma}$ parameters can be calculated fast using just a Euclidean cross product,

i.e. $\hat{\gamma}_j = \mathbf{t}_j^\top \mathbf{y}/||\mathbf{t}_j||_2^2 = \mathbf{t}_j^\top \mathbf{y}/\delta^2{}_j$ for $j = 1, \ldots, r$, where $\mathbf{t}_j$ are columns of $\mathbf{T}$, $||\mathbf{t}_j||_2 = (\sum_i^n |t_{ij}|^2)^{1/2}$ the Euclidean norm and δ_j are the (diagonal) elements of $\mathbf{\Delta}$. Since the $\mathbf{x}$'s are scaled, the constant of the two regression models on $\mathbf{X}$ and on $\mathbf{T}$ are identical and equal to the mean of $\mathbf{y}$ i.e. $\beta_0 = \gamma_0 = \bar{y}$.

Typically one would not regress all columns of $\mathbf{T}$ but only the first λ, i.e. $\mathbf{T}_\lambda$. The manifold $\mathcal{M}(\mathbf{T}_\lambda)$ is the best linear approximation of the original manifold generated by the columns of $\mathbf{X}$, $\mathcal{M}(\mathbf{X})$, in λ-dimensions. The use of PCR this way is claimed to be a computationally efficient model selection technique which also corrects for *multicolinearity*. We will discuss some of the properties of the PCR model below.

3.2 Properties of the PCR model

3.2.1 Model selection technique

Let M denote the rank of the the matrix $\mathbf{X}$. Assuming there are not any pathological co-linearities in $\mathbf{X}$, M will be equal to r if $n > r$ and equal to n if $n < r$. M is the maximum number of scores in the matrix $\mathbf{T}$. Let λ take values in $\{0, 1, \ldots, M\}$. In a typical PCR we choose a specific value of λ and fit only the first λ columns of $\mathbf{T}$, i.e. $\mathbf{T}_\lambda$, and in this case λ plays the role of a tuning (or smoothing) parameter. The case $\lambda = 0$ represents the null model (with only the constant fitted) and $\lambda = M$ represents the full (parameterised) least squares model. Determining which value to choose for λ is a model selection problem. Note that terms with low eigenvalues (the last columns of $\mathbf{T}$) are eliminated from the model. This type of elimination is termed by Hastie et al. (2009), as 'hard-thresholding' compared to 'soft-thresholding' provided by ridge or lasso regression. One great advantage of PCR, (which it shares with ridge and

lasso techniques), compared to a linear model on the original explanatory variables, is the fact that it can work in situations where there are more explanatory variables than observations i.e. when $n \leq r$.

3.2.2 The $\boldsymbol{\beta}$ coefficients

For each tuning parameter value λ there will be different estimated $\boldsymbol{\beta}$ parameters, $\hat{\boldsymbol{\beta}}_\lambda = \mathbf{V}_\lambda \hat{\boldsymbol{\gamma}}_\lambda$ for $\lambda = 1, \ldots, M$. Note that the notation $\mathbf{V}_\lambda$ means the corresponding first λ columns of $\mathbf{V}$, and $\hat{\boldsymbol{\gamma}}_\lambda$ means the first λ values of $\hat{\boldsymbol{\gamma}}$. This series of the estimated $\boldsymbol{\beta}$ parameter can be easily calculated and saved to an $M \times M$ matrix $\mathbf{B}$. Plotting the rows of $\mathbf{B}$ against λ will show the path of how the coefficients $\boldsymbol{\beta}$ change by adding an extra column of the score matrix $\mathbf{T}$ into the model. Those plots, see for example Figure 2, are similar in nature to the ones produced by lasso or ridge regression models when the fitted coefficients are plotted against the tuning parameter, see for example the **glmnet** package of Hastie and Qian (2014).

3.2.3 Variance covariance matrices of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_\lambda$

Because of the orthogonality of the columns of $\mathbf{T}$, the variance covariance matrix for the $\boldsymbol{\gamma}$ coefficients, $\boldsymbol{\Sigma}_\gamma$ is a diagonal matrix. The elements of the diagonal matrix $\boldsymbol{\Sigma}_\gamma$ are $\hat{\sigma}^2_\lambda / \delta^4_\lambda$ for $\lambda = 1, \ldots, M$. The subscript λ in σ is to emphasise that the $\hat{\sigma}^2_\lambda$ is estimated using the residuals from the model using only the first λ columns of $\mathbf{T}$, i.e. $\mathbf{T}_\lambda$. The variance covariance matrix for $\hat{\boldsymbol{\beta}}_\lambda$ is given by $\boldsymbol{\Sigma}_{\beta,\lambda} = \mathbf{V}_\lambda \boldsymbol{\Sigma}_{\gamma,\lambda} \mathbf{V}_\lambda^\top$ where again the subscript λ emphasises that only the first λ columns of the matrices $\mathbf{V}$ and $\boldsymbol{\Sigma}_\gamma$ are used.

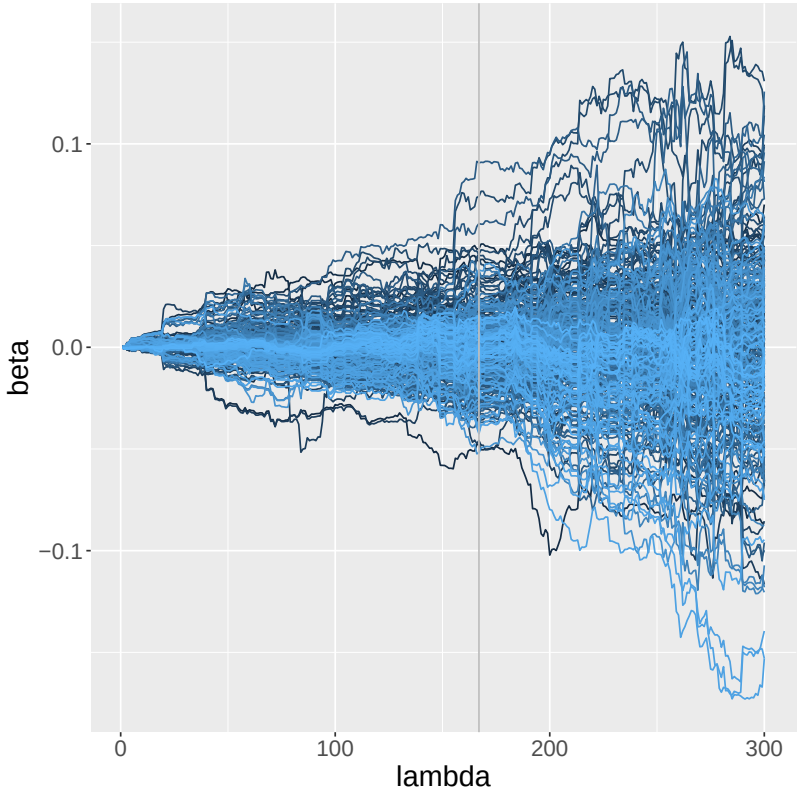

Figure 2: Showing the path of how the $\boldsymbol{\beta}$ coefficients are changing by adding different principal components to the model. The design matrix $\mathbf{X}$ contains all explanatory variables of the Greek spread data plus their first order interactions. Only the first 300 principal components are shown here. The vertical line indicates the number of principal components chosen by using GAIC with $k = \log(1288)$. The model was fitted using the function `fitPCR()` from **gamlss.foreach** package.

3.2.4 Fitted values and residuals

For each tuning parameter value λ there will be different estimated fitted values, $\hat{\mathbf{y}}_\lambda = \mathbf{T}_\lambda \hat{\boldsymbol{\gamma}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$ for $\lambda = 1, \ldots, M$. The residuals also depend on the tuning parameters λ, $\hat{\mathbf{r}} = \mathbf{y} - \hat{\mathbf{y}}_\lambda$ for $\lambda = 1, \ldots, M$.

3.2.5 Estimating λ

For a fixed λ, the observed t-statistic $t_\lambda = \hat{\gamma}_\lambda / \mathtt{se}(\hat{\gamma}_\lambda)$ for $\lambda = 1, \ldots, M$, can be used to access the significance of the coefficient γ_λ, (i.e. by checking whether the observed $t_\lambda > \alpha$, where α is an appropriate value from the tail of the t distribution). Traditionally the value of λ was chosen by including all scores before one of the γ_λ was found to be not significant. That is, if the first coefficient which found to be not significant is $\hat{\gamma}_k$ then the chosen λ is $\hat{\lambda} = k - 1$. This methodology was criticised among others by Jolliffe (1982) and Hadi and Ling (1998). They pointed out that it is very likely that one or more of the components with lower eigenvalues can potentially contribute more in the reduction of the sum of squares of the model than terms with higher eigenvalues. The problem is that while the vectors of scores are ordered (from the highest to the smallest) by having high variances in the linear subspace generated by the rows of $\mathbf{X}$, this does guarantee that those scores also have high correlation with the response. Here are some alternatives method for choosing λ:

GAIC Use an information criterion approach and chose as tuning parameter λ the one which minimises the generalized Akaike information criterion with penalty k. The GAIC is defined as $GAIC(\lambda, k) = -2\ell(\hat{\mu}_\lambda, \hat{\sigma}_\lambda) + k(\lambda + 1)$ where $\ell()$ represents the log-likelihood function of the normal distribution. Note that this

method does not necessarily solve the problem of important explanatory factors with lower eigenvalues.

t-value In a *t-value* approach m scores are fitted, such as $m \leq M$, but only scores with *t-values* greater than, for example, α are included. Note that in this case the tuning parameters are m and α not λ and the 'lower eigenvalues' problem is corrected.

SPCR *Supervised Principal Component Regression* (SPCR) was introduced by Bair et al. (2006). In this approach rather than performing principal component analysis using all the variables, $\mathbf{X}$, we use only a subset of those variables with the strongest estimated correlation with the response say $\mathbf{X_s}$. That is, we first choose the matrix $\mathbf{X_s}$, which is a subset of the original matrix $\mathbf{X}$, containing only columns of $\mathbf{X}$ which have a correlation, in absolute value, with $\mathbf{y}$ higher than say a threshold parameter ρ. The SPCR methodology has two tuning parameters ρ and λ. (In our `R` function `fitPCR()` we fix ρ and estimate λ using GAIC).

PLS *Partial Least Squares*, (PLS), is a technique in which the orthogonal decomposition of the design matrix $\mathbf{X}$ is done in such a way that the orthogonal components with sequentially the highest correlation to the response variable are chosen, see Wold (1975) Hastie et al. (2009) pp 80, Wehrens and Mevik (2007). In practice it is found that while PLS reduces the degrees of freedom of the fitted model, it does not necessarily perform better than PCR, see Wentzell and Vega-Montoto (2003). Also the fit is more computationally demanding. We will not pursue this method in this paper.

3.2.6 Multicolinearity

Multicollinearity is defined as the problem of having highly correlated linear terms in the model. High correlations between the explanatory variables result in unstable fitted linear coefficients which makes the interpretation of the coefficients problematic. The columns of the scores $\mathbf{T}$ (the ones with non-zero eigenvalues) are orthogonal and therefore the parameters $\boldsymbol{\gamma}$ do not suffer from multicollinearity. Interpretation of the model via the $\boldsymbol{\beta}$ coefficients could however be more problematic. Artigue and Smith (2019) claimed that the 'estimated coefficients are distorted by PCR in ways that diminish the accuracy of the model when it is used to make predictions with fresh data'. Note that there are other techniques in the literature (like lasso or elastic net) which can correct for multicollinearity.

3.2.7 Prior weights and the function `fitPCR()`

We have implemented the simple PCR in **R** in the function `fitPCR()` within the package **gamlss.foreach**. This function is very similar to the function `svdpc.fit()` of the package **pls** in CRAN but with the additional feature of prior weights. Prior weights are needed for a GAMLSS implementation of PCR. The prior weights were implemented by: (a) scaling $\mathbf{X}$ using *weighted* means and standard deviations. (b) transforming $\mathbf{y}$ and $\mathbf{X}$ to $\mathbf{y_w} = \sqrt{\mathbf{w}} \circ \mathbf{y}$ and $\mathbf{X}_w = \sqrt{\mathbf{w}} \circ \mathbf{X}$, respectively, (where $\circ$ symbolise the Hadamard element by element product) and finally (c) taking the singular value decomposition of $\mathbf{X}_w$. The function `fitPCR()` is one of the two methods we used to implement PCR in GAMLSS, the subject of the next section.

3.3 The GAMLSS algorithms for PCR

There are two implementations of PCR within the GAMLSS framework. They differ in the time they take to perform the singular value decomposition (s.v.d) within the GAMLSS fitting algorithm. The GAMLSS algorithm is described in detail on Chapter 3 of Stasinopoulos et al. (2017). [The algorithm requires, at each iteration, a working (response) variable $\mathbf{y}^{(i)}$ and working vector of weights $\mathbf{w}^{(i)}$, which are both functions of the first and second derivatives of the log-likelihood with respect to the appropriate distribution parameters.] In the first implementation, the function `pc()`, performs the s.v.d. on $\mathbf{X}$ at the beginning of the fitting algorithm as described in the Algorithm 1. In the second method, the function `pcr()`, performs s.v.d. on $\mathbf{X_w}$ each time within the backfitting algorithm of GAMLSS, see Algorithm 2. Note that, in Algorithm 1, the weighted column vectors of the scores of $\mathbf{X}$ are not orthogonal and therefore, estimating the $\boldsymbol{\gamma}$'s, at stage 4, requires a proper weighted least squares fit. In contrast in Algorithm 2, the $\boldsymbol{\gamma}$'s are calculated quickly using crossproducts but the recalculation of the s.v.d. of $\mathbf{X_w}$ each time slows down the performance and introduces extra instability in the algorithm. Both Algorithms 1 and 2 show the case in which the estimation of λ is achieved using GAIC. The algorithms would have to be amended for the t-values approach. Supervised PCR can be done before the start of Algorithms 1 and 2.

4 Results

We have found PCR very useful in modelling the Greek spreads because it allowed first order interactions between the 67 financial indicators to be modelled. At an early

Algorithm 1 : perform s.v.d. before iterative $\mathbf{y}^{(j)}$ and $\mathbf{w}^{(j)}$ are defined

scale $\mathbf{X}$ and evaluate $\mathbf{X} = \mathbf{UDU}^\top = \mathbf{TP}$

for $\mathbf{y}^{(j)}$ and $\mathbf{w}^{(j)}$ until convergence **do**

  for $\lambda = 1, 2, \ldots, M$ **do** regress $\mathbf{y}^{(j)}$ on $\mathbf{T}_\lambda$ with weights $\mathbf{w}^{(j)}$, i.e.

    calculate $\hat{\boldsymbol{\gamma}}_\lambda = [\mathbf{T}_\lambda \mathbf{W} \mathbf{T}_\lambda]^{-1} \mathbf{T}_\lambda \mathbf{W} \mathbf{y}^{(j)}$

    compute fitted values $\hat{\mathbf{y}}_\lambda^{(j)} = \mathbf{T}_\lambda \hat{\boldsymbol{\gamma}}_\lambda$

    get the local $GAIC_\lambda = \sum_i^n w_i (y_i - \hat{y}_{i,\lambda})^2 + k * df_\lambda$

  end for

  The $\hat{\lambda}$ that corresponds to the minimum GAIC is chosen

  recalculate $\hat{\boldsymbol{\gamma}}_{\hat{\lambda}} = [\mathbf{T}_{\hat{\lambda}} \mathbf{W} \mathbf{T}_{\hat{\lambda}}]^{-1} \mathbf{T}_{\hat{\lambda}} \mathbf{W} \mathbf{y}^{(j)}$ and $\hat{\mathbf{y}}_{\hat{\lambda}}^{(j)} = \mathbf{T}_{\hat{\lambda}} \hat{\boldsymbol{\gamma}}$

  calculate $\hat{\boldsymbol{\beta}}_{\hat{\lambda}} = \mathbf{V}_{\hat{\lambda}} \hat{\boldsymbol{\gamma}}_{\hat{\lambda}}$

end for

Algorithm 2 : s.v.d. after iterative $\mathbf{y}$ and $\mathbf{w}$ are defined (with local GAIC)

for $\mathbf{y}^{(j)}$ and $\mathbf{w}^{(j)}$ until convergence **do**

  evaluate $\mathbf{X}_w = \sqrt{\mathbf{w}} \circ \mathbf{X} = \mathbf{U}_w \mathbf{D}_w \mathbf{U}_w^\top = \mathbf{T}_w \mathbf{P}_w$ and $\mathbf{y}_w^{(j)} = \sqrt{\mathbf{w}} \circ \mathbf{y}^{(j)}$

  get $\hat{\boldsymbol{\gamma}}_\lambda = \mathbf{t}_\lambda^\top \mathbf{y}_w^{(j)} / \delta_{w,\lambda}$ $\lambda = 1, 2, \ldots, M$ (The columns of $\mathbf{T}_w$ are orthogonal).

  for $\lambda = 1, 2, \ldots, M$ **do**

    get $\hat{\boldsymbol{\beta}}_\lambda = \mathbf{V}_\lambda \hat{\boldsymbol{\gamma}}_\lambda$

    get fitted values and residuals $\hat{\mathbf{y}}_\lambda^{(j)} = \mathbf{X}_w \hat{\boldsymbol{\beta}}_\lambda$, $\hat{\mathbf{r}}_\lambda^{(j)} = \mathbf{y}^{(j)} - \hat{\mathbf{y}}_\lambda^{(j)}$

    get $\hat{\sigma}_\lambda = \sum_i^n \left[\hat{\mathbf{r}}_\lambda^{(j)}\right]^2 / n$

  end for

  get local $GAIC(\lambda, k) = -2 \sum_i^n \log \texttt{NO}\left(y_i, \hat{\mu}_\lambda, \hat{\sigma}_\lambda\right) + k(\lambda + 1)$ for $\lambda = 1, 2, \ldots, M$ and the $\hat{\lambda}$ corresponds to the minimum GAIC is selected

end for

stage of the analysis, it became clear that including all the 67 financial indicators as linear or in fact as non-linear smoothing terms (in the models for μ and σ) failed to account properly for the actual trend in spreads. The inclusion of first order interactions improved the model considerably. This is consistent with the *sparsity of effects principle* which states that 'a system is usually dominated by main effects and low-order interactions', Surhone et al. (2011). However if there are r different linear (continuous) terms in a design matrix, the total number of columns including main and first order interactions is $r(r+1)/2$. In our case, we have 1288 cases and 67 explanatory variables therefore the design matrix will have $(67 \times 68)/2 = 2278$ columns and clearly there are more variables than observations.

4.1 Normal distribution

The analysis started by assuming a normal distribution for the response variable. Table 1 shows 18 different fitted models all using the normal distribution for the response variable (Greek spreads). Models 1 to 5 and models 11 to 14 are models where only the mean, μ, is modelled, while model 6 to 10 and models 15 to 18 have both the mean, μ, and the standard deviation, σ, modelled using explanatory variables. Models 1 to 10 used the main effects of the 67 explanatory variables while models 11 to 18 used the main effects plus the first order interactions. The two different algorithms used, as described in Algorithms 1 and 2, are denoted in the table as 'pc' and 'pcr', names corresponding to their `R` functions `pc()` and `pcr()`. Also the two methods for the selection of the tuning parameter λ , GAIC and 't-values' are shown in Table 1 as 'gaic' and 't-val', respectively.

By using AIC (or BIC) modelling simultaneously μ and σ proved to be superior to

modelling just only the μ. The best model overall proves to be model 15 where the Algorithm 1 was used and where the number of principal components was chosen using the "t-values" approach. Figure 3(a) shows the data, the fitted values and the residuals (lower part of the plot) from model 7, the best model without interactions. Figure 3(b) shows the data, the fitted values and residuals from model 15 the best model with first order interactions. It is apparent that the interaction model 15 managed to capture the trend of the Greek spreads well, while model 7 failed to do so. Figure 3(c) and (d) show the fitted σ's from model 7 and 15 respectively. Note that in the fitted values for σ, at model 15, they are two days (23 and 24 of March 2009) where the predicted σ's have large values (> 1), but in which the observed spreads have values relatively close to the fitted μ and therefore giving relatively small residuals.

4.2 Different distributions

Next, using model 15 as a basis, we tried fitting different distributions. Figure 4 shows the *ordered* AIC for the different fitted distributions. The x-axis is scaled from 0 to 1, (zero for the 'worst' fitted model and one for the 'best'). The ordering of the distributions is done using the formula, $O_D = (AIC_{max} - AIC_D)/(AIC_{max} - AIC_{min})$, where AIC_D is the GAIC of distribution D, and AIC_{min} and AIC_{max} are the AIC's of the best and worst fitted distribution, respectively. Hence the length of the bar of a distribution in the plot indicates its AIC compared to the two extremes. Rigby et al. (2019) provides more information about the different distributions fitted and the notation used in the plot. The worst fitting distribution for the Greek spreads is `WEI3` (Weibull parametrisation type 3, where μ is the mean of the distribution), while

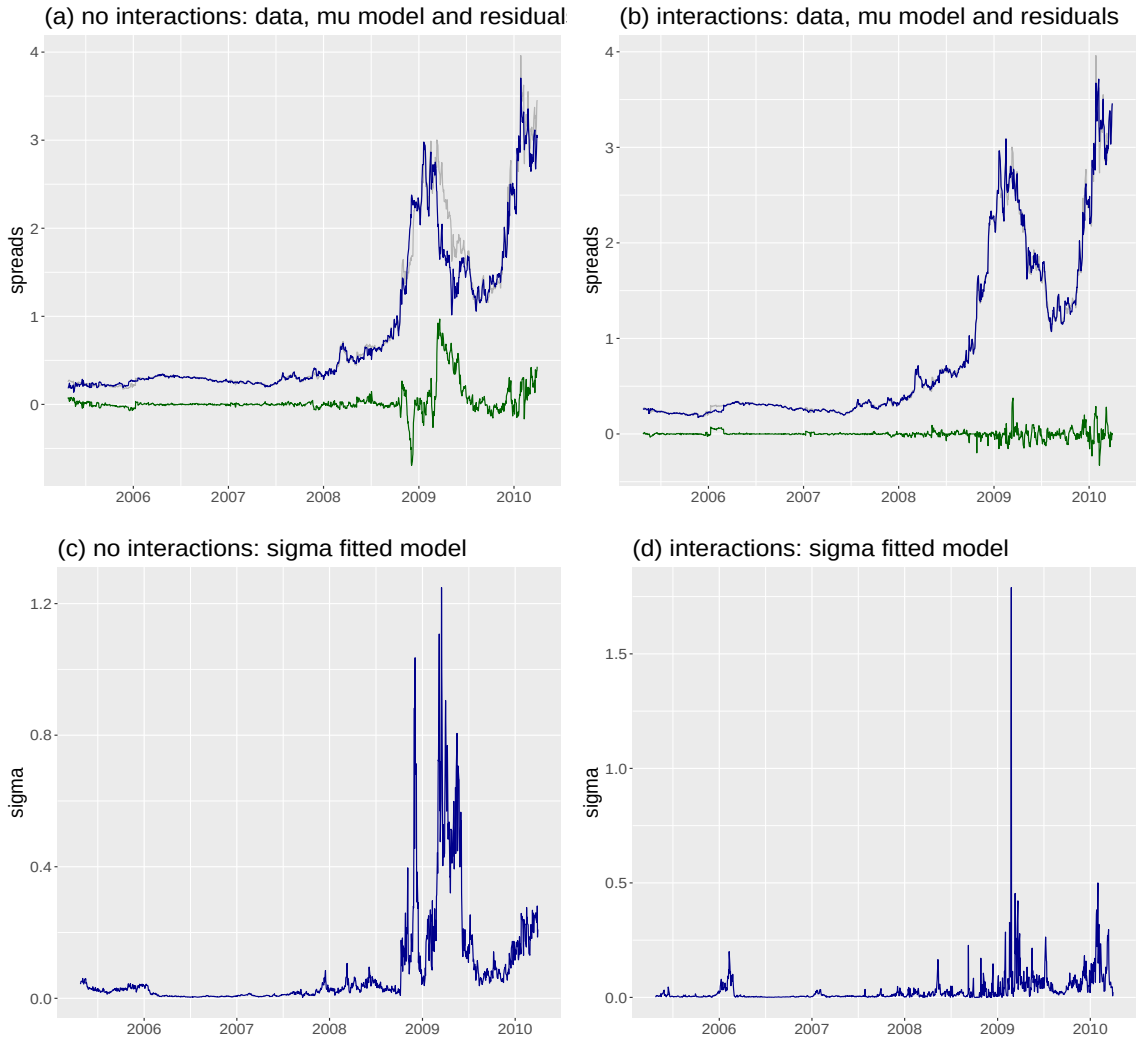

Figure 3: Showing results for the normal distribution analysis. Both top panels (a) and (b) show the actual data (Greek spreads during the period from 25th of April 2005 to the 31th of March 2010) and the fitted values from the best model with no interactions, (model 7), on the left, and the best model with interactions, (model 16), on the right. The curves below the data show the difference between the actual data and the fitted values for the mean model. Panels (c) and (d) show the fitted values for σ for the best models without and with interactions, i.e. models 7 and 16, respectively.

the Box-Cox t-distribution, `BCT`, comes as best. The `BCT` distribution, Rigby and Stasinopoulos (2006), is a four parameter distribution with μ, a location parameter, approximately the median, σ approximately the coefficient of variation, and ν and τ representing skewness and kurtosis parameters, respectively. Note that all the parameters of the distributions displayed in Figure 4 were fitted using a PCR model. Trying to simplify the `BCT` model, we have refit it, firstly, with parameter τ as a constant and secondly, with both ν and τ parameters as constants. The BCT model, with all parameters fitted as PCR produced better AIC and BIC, using 190 parameters. We decided to select the model with both ν and τ as constant because it displayed as good residuals and had fewer fitted parameters, 158. Figure 5(a) shows the worm plot, van Buuren and Fredriks (2001), of the residuals from the normal (model 15 of Table 1), Figure 5(b) from the `BCT` model with PCR for all the parameters, Figure 5(c) from the `BCT` with τ as a constant, and finally Figure 5(d) shows the worm plot from the `BCT` model, with both ν and τ as constants. In Figure 5(d) all points of the worm plot were within the point-wise acceptance region of the worm plot. This model had 106 principal components for the μ model and 46 for the σ model with ν and τ constants.

4.3 Interpretation of the results

The BCT model can be simplified further by noting that the fitted parameter for τ is rather large and that as $\tau \to \infty$ the BCT distribution becomes the BCCG distribution. The Box, Cox, Cole and Green (BCCG) distribution, has three parameters and it is generated similarly to the BCT distribution but its modified Box-Cox transformation assumes that the original variable is normal rather than t distributed,

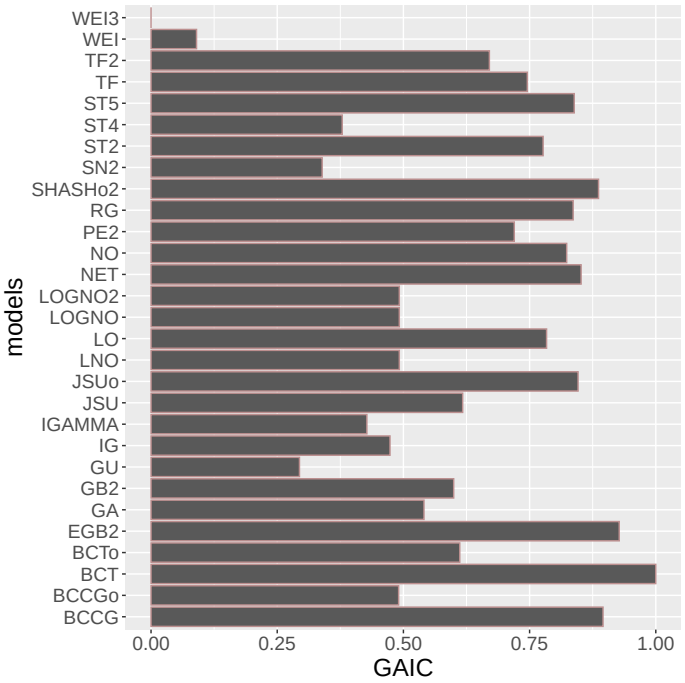

Figure 4: Plot showing the ordering, from zero (worst) to one (best), in terms of AIC of the different distributions fitted to Greek spreads data. The Box-Cox t distribution is best.

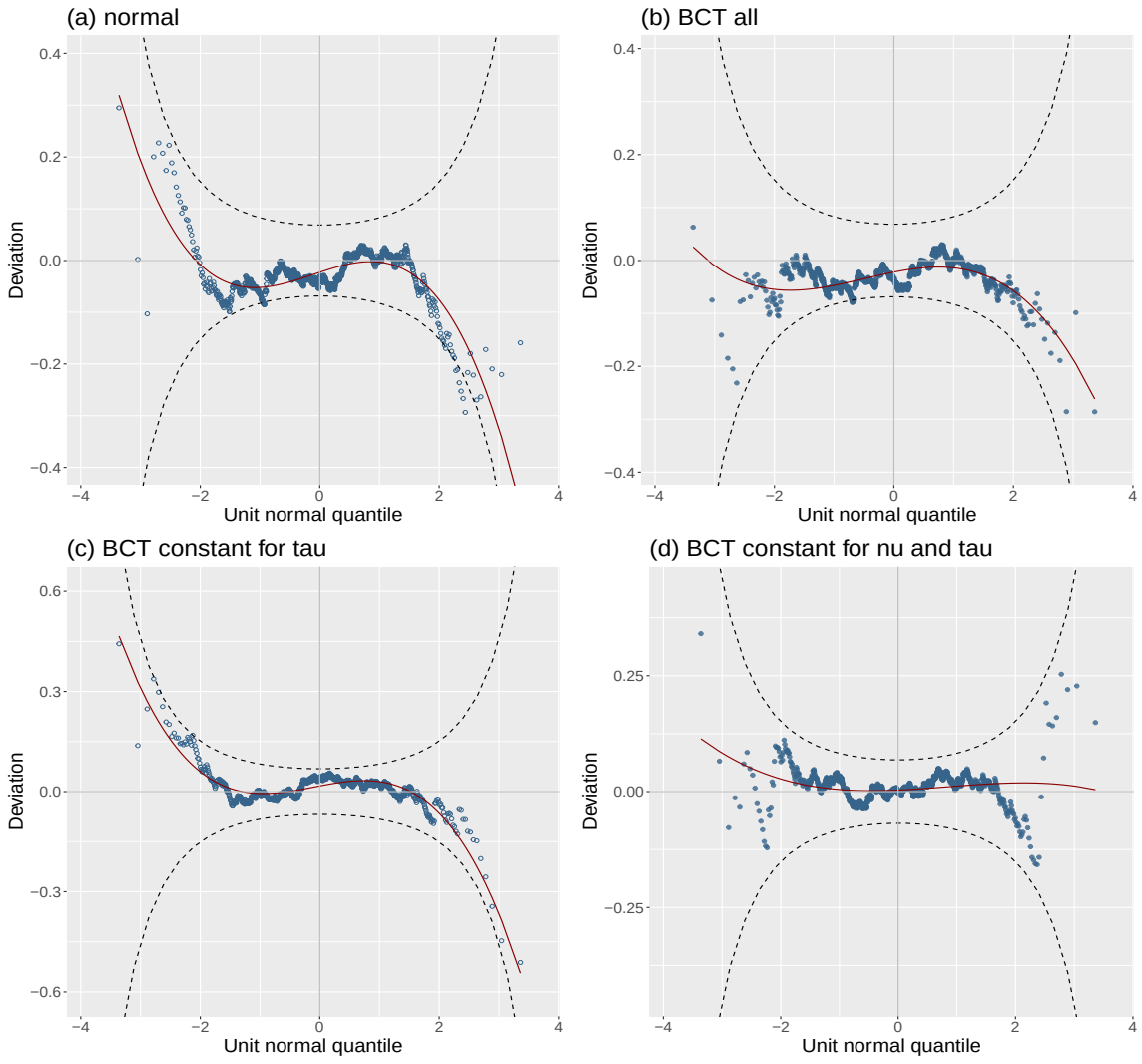

Figure 5: Showing worm plots from (a) the normal distribution model 15 from Table (1), (b) from the BCT model with all parameters modelled with a PCR model, (c) BCT with parameters τ as a constant, (d) BCT with both ν and τ as constants.

respectively, [see Section 19.4.1 in Rigby et al. (2019)]. The re-fitted BCCG distribution model with PCR model for μ and σ but not for ν had also a smaller GAIC value than the equivalent BCT model, i.e. -7153.21 with 158 degrees of freedom compared to -7135.03 with 154 degrees of freedom. The number of degrees of freedom for the PCR μ model was decreased by 1 while for the σ model this was increased by 6. In addition the worm plot from the BCCG model (not shown here) appears as good as the one from the BCT model shown in Figure 5(d). As a consequence the BCCG distribution model is adopted as the final model:

$$\begin{aligned} \mathbf{y} &\overset{\text{ind}}{\sim} \texttt{BCCG}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \\ g_1(\boldsymbol{\mu}) &= \mathbf{T}_{(1,105)}\boldsymbol{\gamma}_1 \\ g_2(\boldsymbol{\sigma}) &= \mathbf{T}_{(2,52)}\boldsymbol{\gamma}_2 \\ g_3(\boldsymbol{\nu}) &= \beta_{30} = -2.317 \end{aligned} \tag{4.1}$$

The fitted value for the skewness parameter, ν, is $\hat{\nu} = -2.317$ with a standard error 0.579 indicating a right skew distribution for the spreads. Since a BCCG distribution with $\nu = 0$ is the log-normal, we can conclude at this point, that the log-normal distribution is not supported here. The shape of the fitted distribution changes dramatically at different periods of time. Figure 6 demonstrated the different shapes of the fitted `BCCG` distribution split into four time intervals. Figure 6(a) shows the different shapes of fitted distributions from the 25th of April 2005 to the 18th of July 2006. Note that the distributions shown are separated by 7-day intervals. Figure 6(b) shows the fitted distribution from the 19th of July 2006 to the 11th of October 2007. Figure 6(c) shows the fitted distribution from the 14th of October 2007 to the 5th of January 2009. Finally figure 6(d) hows the fitted distribution from the 6th of January 2009 to the 31th of March 2010. Note how the quantile range of the fitted

distributions of the response is changing over time reflecting both the decrease in value but also the volatility of the Greek spreads.

The interpretation of the PCR models for μ and σ in (4.1) is rather challenging because of the large number of coefficients involved. For example, in order to interpret the model for μ, we have to examine the corresponding 2278 $\boldsymbol{\beta}$ coefficients rather than the 105 fitted $\hat{\boldsymbol{\gamma}}$'s, because it is the $\boldsymbol{\beta}$ coefficients relating the explanatory variables and their interactions to μ. Rotating the 105 PC, a trick often done with a smaller number of PCs, see Jolliffe (2002), is also unlikely to help the interpretation. The corresponding fitted $\hat{\boldsymbol{\beta}}$ coefficients for the μ and the σ models are shown in Figure 7 (a) and (b), respectively. Our strategy is to choose a cut off point and identify which $\hat{\boldsymbol{\beta}}$ coefficients have an absolute value greater than the cut off point. Since the design matrix is standardised this should point us to the most influential factors in determine μ and σ. For the μ model we have chosen a cutoff point of 0.04, identifying 27 different terms, while for the σ model we have chosen a cutoff point of 0.45 resulting to 10 different terms. Those cutoff values are shown as vertical lines in Figures 7(a) and (b) respectively. For the μ model all 27 influential terms are one way interactions. Prominently featuring are interactions including:

country risk factors: Greek, French, Italian and German 10 year bond spreads, Euro Generic Govt Bond 10 Year, ,

liquidity factors: like: FTSE Euro Corporate Bonds, British Banker Association (BBA) index, EONIA overnight index average, benchmark rate, JP Morgan EMBI index , Credit Default Swap (CDS), US Generic Govt 10 Year Yield

volatility factors Dow Jones EURo stock, Chicago Board Options Exchange,

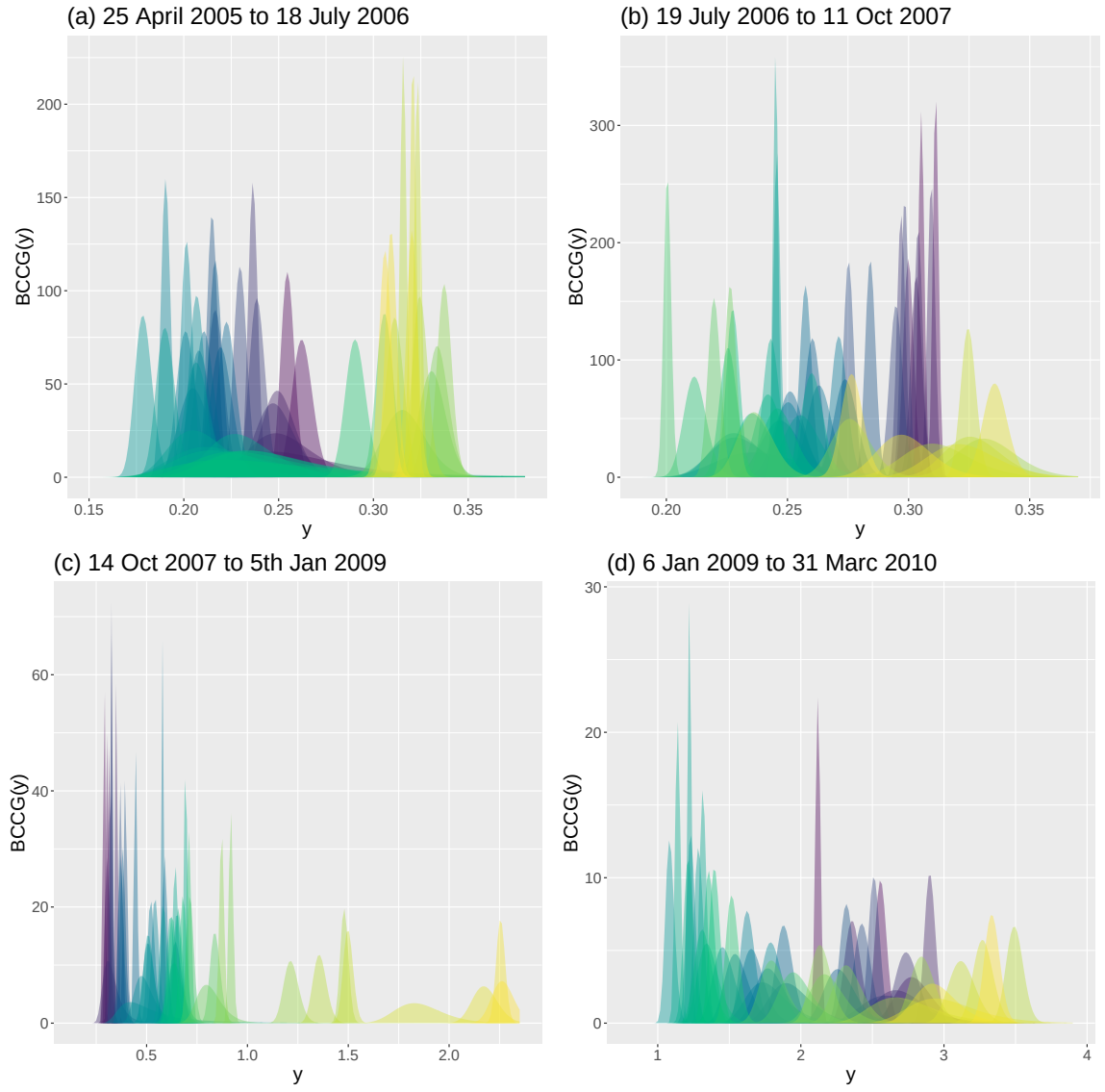

Figure 6: Showing the fitted distributions of the `BCCT` model (at an interval of 7 days): (a) from the 29th of April 2005 to the 18th of July 2006 (b) from the 19th of July 2007 to 11th of October 2007, (c) from the 14th of October 2007 to the 5th of January 2009 and (d) from the 6th of January 2009 to the 31 of March 2010. Note the different ranges of the response variable over time.

Stock exchange indexes Euro Stock 50 price, Austrian Trade Index, Benchmark Stock Market Index of Euronext, Paris Bourse stock exchange, DAX, Amsterdam Exchange Index

For the σ model there is one main effect of the Netherlands 10 year bond spread (country risk). The interactions appearing in σ are coming from: FTSE Euro Corporate Bonds (Liquidity), Greece Index (Country Risk), Euro Generic Government Bond 2 Year (Country Risk), Austrian Trade Index, Portuguese Republic index (country risk), Kingdom of Spain Index 10 Year (country risk), Chicago Board Options Exchange (Volatility), Credit Default Swap (CDS) Index (Liquidity),

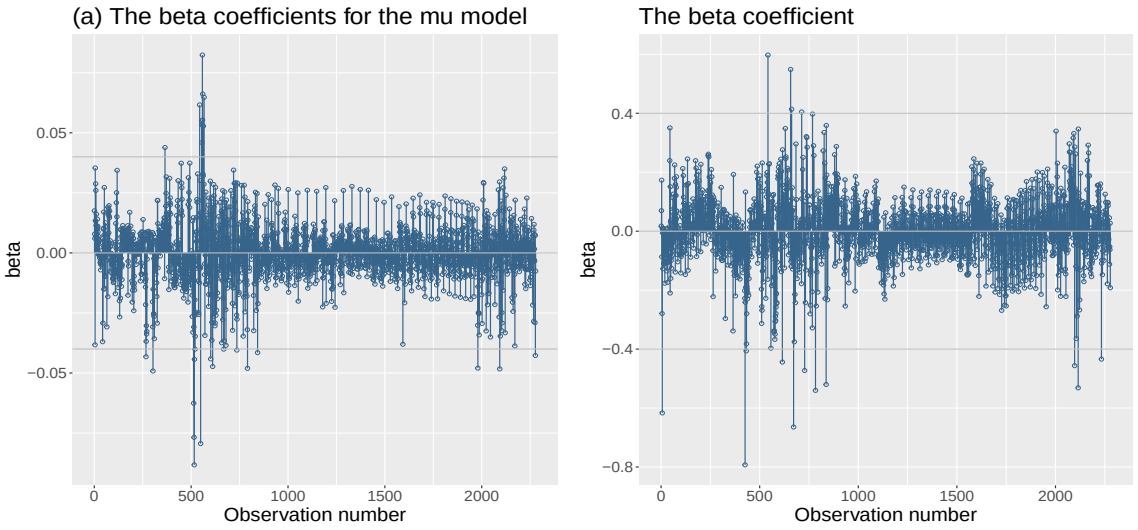

Figure 7: Showing the fitted beta coefficients from (a) the μ model and (b) for the σ model.

5 Conclusions

We have shown that by including first order interactions of continuous explanatory variables and by using principal component regression for all the distribution param-

eters of the response we were able to model rather complex economic relationships. The methodology in doing so is rather general and can be applied in a variety of data sets not necessary economic in nature.

The following comments are important here. The scope for the original data collection was to explore empirically the relationship of the Greek spreads against relevant economic indicators. Unfortunately the GAMLSS-PCR approach, used here, has rather complex intepretation. If the scope of the analysis is prediction, then a more dynamic model, including lags for the explanatory variables and possibly also of the response, is needed. We will explore that in future work. Also because of the time series nature of the data a small time series component (i.e. autocorrelation) remains in the residuals even after fitting μ and σ. This requires further investigation.

To conclude we would like to mention a story imprinted in the memory of the first author of this paper encountered while he was working in the Centre of Applied Statistics at University of Lancaster under Murray Aitkin in the nineteen-eighties. On this occasion both Murray and he were walking together towards the printer room to collect computer output. On arrival, when Murray saw the big pile of printout waiting for him, he cried "Oh no, I forgot to put the convergence criterion in the macro". It was a GLIM macro (he still uses GLIM today) and he was testing his mean and variance modelling idea which later became his Aitkin (1987) paper. It was this basically simple idea of recursively fitting the mean and variance which led some years later to the creation of GAMLSS. We are indebted to him. Thank you, Murray for your kindness and your openness to share ideas. We are also looking forward to your ninetieth birthday celebrations.

6 Acknowledgments

The authors would like to thank the referees for their helpful comments and Dr Tilemachos Efthimiadis who help to collect the data. De Bastiani wishes to acknowledge CNPq, Process number 310050/2019-7.

References

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Applied Statistics*, **36**, 332–339.

Artigue, H. and Smith, G. (2019). The principal problem with principal components regression. *Cogent Mathematics & Statistics*, (just-accepted), 1622190.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101**(473), 119–137.

Hadi, A. S. and Ling, R. F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, **52**(1), 15–19.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2nd edition.

Hastie, T. and Qian, J. (2014). Glmnet vignette. *Retrieve from http://www. web. stanford. edu/˜ hastie/Papers/Glmnet_Vignette. pdf. Accessed September*, **20**, 2016.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, **10**(2), 69–79.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **31**(3), 300–303.

Jolliffe, I. (2002). Rotation and interpretation of principal components. *Principal Component Analysis*, pages 269–298.

Kendall, M. G. (1957). *A course in Multivariate Analysis*, volume 620. Griffin, London.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics*, **54**, 507–554.

Rigby, R. A. and Stasinopoulos, D. M. (2006). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, **6**(3), 209.

Rigby, R. A., Stasinopoulos, D. M., Heller, G. Z., and De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. Chapman & Hall/CRC, Boca Raton.

Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman & Hall/CRC, Boca Raton.

Stasinopoulos, D. M., Rigby, R. A., and De Bastiani, F. (2018). GAMLSS: a distributional regression approach. *Statistical Modelling*, **18**(3-4), 248–273.

Surhone, L., Tennoe, M., and Henssonow, S. (2011). *Sparsity-Of-Effects Principle*. Betascript Publishing. ISBN 9786136331782. URL https://books.google.com.br/books?id=IIrlmQEACAAJ.

van Buuren, S. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259–1277.

Wehrens, R. and Mevik, B.-H. (2007). The pls package: principal component and partial least squares regression in r. *Retrieve from https://CRAN.R-project.org/package=pls*.

Wentzell, P. and Vega-Montoto, L. (2003). Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, **65**, 257–279. doi: 10.1016/S0169-7439(02)00138-7.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, **12**(S1), 117–142.

Table 1: A table of deviance, AIC and BIC from the fitted normal distribution models. The notation "pc" and "pcr" refer to algorithms 1 and 2, respectively, and the corresponding `R` functions `pc()` and `pcr()`.

Models	df	Deviance	AIC	BIC
1. linear, μ	69.00	-2727.75	-2589.75	-2233.65
2. pc, t-val, μ	48.00	-2700.59	-2604.59	-2356.87
3. pcr, t-val, μ	69.00	-2727.75	-2589.75	-2233.65
4. pc, gaic, μ	52.00	-2673.38	-2569.38	-2301.02
5. pcr, gaic, μ	52.00	-2673.38	-2569.38	-2301.02
6. linear, μ, σ	136.00	-5441.82	-5169.82	-4467.95
7. pc, t-val, μ, σ	77.00	-5341.63	**-5187.63**	**-4790.24**
8. pcr, t-val, μ, σ	106.00	-5392.65	-5180.65	-4633.60
9. pc, gaic, μ, σ	82.00	-5227.46	-5063.46	-4640.27
10. pcr, gaic, μ, σ	80.00	-5224.72	-5064.72	-4651.85
11. pc, t-val, μ, inter.	139.00	-5319.57	-5041.57	-4324.21
12. pcr, t-val, μ, inter.	201.00	-5411.75	-5009.75	-3972.42
13. pc, gaic, μ, inter.	169.00	-5240.37	-4902.37	-4030.19
14. pcr, gaic, μ, inter.	169.00	-5240.37	-4902.37	-4030.19
15. pc, t-val, μ, σ, inter.	221.00	-8217.34	**-7775.34**	**-6634.79**
16. pcr, t-val, μ, σ, inter.	276.00	-8152.55	-7600.55	-6176.16
17. pc, gaic, μ, σ, inter.	185.00	-7042.32	-6672.32	-5717.57
18. pcr, gaic, μ, σ, inter.	174.00	-7075.67	-6727.67	-5829.68