qgam: computer lab exercises

Please make sure that the following packages are installed:

1. `devtools` (`install.packages("devtools")`);
2. `qgam` (`library(devtools); install_github("mfasiolo/qgam")`);
3. `mgcViz` (`library(devtools); install_github("mfasiolo/mgcViz")`);
4. `gamair` (`install.packages("gamair")`).

The exercises you could try in the morning session are:

1. Bone mineral density modelling;
2. Self-paced reading latencies for Russian;

while in the afternoon session you could try:

1. CO_2 modelling;
2. Electricity load forecasting;
3. Reaction times for Estonian case-inflected nouns;
4. Rainfall modelling in Switzerland;

but feel free to try `qgam` on your own data.

1 Bone mineral density modelling

This dataset is taken from the package `lava`. It consists of 112 girls randomized to receive calcium or placebo. The response variable of interests consists of longitudinal measurements of bone mineral density (g/cm^2) measured approximately every 6th month for 3 years. All girls are approximately 11yo at the start of the trial. The main variables are:

- `bmd` bone mass density;
- `group` placebo or supplement;
- `person` factor indicating the id of each girl;
- `age` the age of each girl at the time of each measurement;

Questions:

1. Load `mgcViz` and the data with `load("data/calcium.rda")`. Then use `qgamV` to fit a median model with `bmd` as response and linear effects for `age` and `group`. Use `summary` to print the model output. Is the placebo effect significant? (which is the same as asking whether the treatment effect is significant)
2. Use `check1D` with the `l_gridQCheck1D` layer to check that the proportion of negative residuals does not depart too much from 0.5, for any of the subjects. If you see significant departures add a random effect for `person` to the models formula, then re-fit and re-check the residuals. Print the model output again using `summary`.

3. Now modify the model formula to use a smooth effect for `age`, and plot the fitted effects using `plot`. Use the function `AIC` to compare the model with a smooth effects for `age` with the model which uses a linear `age` effect. Which model achieves lower AIC?

4. Verify whether the smooth age effect is different between the placebo and the treatment group, by using a by-factor smooth. To do this substitute `s(age)` with `s(age, by=group)` in the model formula, refit and then plot the fitted effects. To see the difference between the two smooths more clearly, use the `plotDiff` function with the `l_fitLine` and `l_ciLine` layers.

5. Use `mqgam` to fit the same model to the quantiles `qu = seq(0.2, 0.8, length.out = 5)`, and plot the fitted models using `plot` with argument `allTerms = TRUE` (to plot also the parametric effects).

2 Self-paced reading latencies for Russian

Here we look at how the reaction time is influenced by two neural network learning measures (ActOrtho and ActTAM) and possibly subjects' learning rate (SubjectSpeedUp). Position in sentence, trial and subject ID are control predictors. The main variables are:

- RT: reaction time in a self-paced reading task
- ActOrtho: support from orthographic visual input
- ActTAM: support from Tense/Aspect/Mood
- SubjectSpeedUp: subject measure from separate task
- PositionInSentence: position of the verb in the sentence
- Subject: identifier of participant in experiment
- Verb: the "try" verb on which reaction time is measured
- Trial: the (scaled) trial number in the experiment

Questions:

1. Load the `mgcViz` R package and the data with `load("data/russian.rda")`. Create a model formula with `RT` as response and linears effects for `ActOrtho`, `ActTAM`, `SubjectSpeedUp`, `PositionInSentence` and `Trial`.

2. Fit a quantile GAM model for the median using the `qgamV` function. Use `summary` to verify the magnitude and significance of each linear effect.

3. Let `fit` be the output of the last call to `qgamV`. Call `check1D(fit, x = russian$Subject)` with the `l_gridQCheck1D` layer to verify how the proportion of observations falling below the fit changes depending on the subject considered. Given that we are estimating the median, we expect the proportion to be close to 0.5. Do you see significant departures from 0.5 and, if so, what effect would you include to correct this problem?

4. Modify the model formula to include a random effect per subject (`s(Subject, bs = "re")`), and refit the model. Use again `check1D` to verify whether the residual pattern is gone. Also compare the model with and without a random effect for subject using the Akaike Information Criterion (function `AIC`). Does the larger model achieve a lower AIC? Use `summary` to verify whether the effects of `ActOrtho` and `ActTAM` are now significant.

5. Now use the `mqgamV` function to fit this model to the quantiles `qus = seq(0.1, 0.9, length.out = 11)`. If the output of `mqgamV` is called `fit`, plot all the fitted effects using `plot(fit, allTerms = TRUE)`. You should see that the effect of `PositionInSentence` varies greatly between quantiles, can you provide an explanation regarding why this could be the case (there is no right answer)?

6. Modify the model formula by substituting all the linear effects with smooth effects (eg. `s(ActTAM)`) and refit using `mqgamV`. Compare the AIC of the model you just fitted with the simpler model fitted in the previous point. If you plot all the effects as before, you will see large difference between quantiles. However, some of these are probably spurious given that the p-values of the fitted effects are fairly large (if `fit` is the output of `mqgamV`, then `summary(fit[[6]])` shows the coefficient and p-values of the QGAM model for the sixth quantile).

3 CO_2 modelling

This question[1] is about modelling data with seasonality, and the need to be very careful if trying to extrapolate with QGAMs (or any statistical model). The data frame `co2s` contains monthly measurements of CO_2 at the south pole from January 1957 onwards. The columns are `co2`, the month of the year, `month`, and the cumulative number of months since January 1957, `c.month`. There are missing `co2` observations in some months.
Questions:

1. Load `mgcViz` and the data with `library(gamair); data(co2s)`

2. Plot the CO_2 observations against cumulative months.

3. Fit an additive model for the median, $\mu_{0.5}(\texttt{c.month}_i) = f(\texttt{c.month}_i)$ where f is a smooth function, using the `qgam` function. Use the `cr` basis, and a basis dimension of 100. Set $\texttt{err} = 0.1$ in `qgam` to avoid warnings and speed up computation.

4. Obtain the predicted CO_2 for each month of the data, plus 36 months after the end of the data, as well as associated standard errors. Produce a plot of the predictions with twice standard error bands. Are the predictions in the last 36 months credible? NB: to produce the plot you have to write your own code, `mgcViz` does not produce such plots.

5. Fit the model $\mu_{0.5}(\texttt{c.month}_i) = f_1(\texttt{c.month}_i) + f_2(\texttt{month}_i)$ where f_1 and f_2 are smooth functions. Use a basis of dimension 50 for f_1 and a cyclic basis for f_2. In the `qgam` call, you will need to set argument `argGam` to `list(knots=list(month=c(1,13)))` to make so that that the effect of January is the same as January, not that December and January are the same!

6. Repeat the prediction and plotting in question 4 for the new model. Are the predictions more credible now? Explain the differences between the new results and those from question 4.

4 Electricity load forecasting

Here we consider a UK electricity demand dataset, taken from the national grid. The dataset covers the period January 2011 to June 2016 and it contains the following variables:

- `NetDemand` net electricity demand between 11:30am and 12am.

[1] Shamelessly adapted from Simon Wood's notes

- wM instantaneous temperature, averaged over several English cities.
- wM_s95 exponential smooth of wM, that is wM_s95[i] = a*wM[i] + (1-a)*wM_s95[i] with a=0.95.
- Posan periodic index in [0, 1] indicating the position along the year.
- Dow factor variable indicating the day of the week.
- Trend progressive counter, useful for defining the long term trend.
- NetDemand.48 lagged version of NetDemand, that is NetDemand.48[i] = NetDemand[i-2].
- Holy binary variable indicating holidays.
- Year and Date should obvious, and partially redundant.

Questions:

1. Load mgcViz and the data (data("UKload")). Then create a model formula (e.g. y~s(x)) containing: smooth effects for wM, wM_s95, Posan and Trend with 20, 20, 4 and 50 knots and cubic regression splines bases (bs='cr'); parametric effects for Dow, NetDemand.48 and Holy.

2. Use the qgamV function to fit this model for the median, setting err=0.1 to avoid numerical problems. Call fit the fitted model and use plot(fit) and summary(fit) to visualise the fitted effects and to see which effects are significant. Do you notice anything problematic about the effect of Posan? How many degrees of freedom are we using for this smooth effect (you can read it from the output of summary)?

3. Modify the effect of Posan to use an adaptive (bs='ad') spline basis. Then refit the model and plot the smooth effects. Has the effect of Posan changed? How many degrees of freedom are we using now for Posan? Explain what happened.

4. Use mqgamV to fit this model to the five quantiles qu=seq(0.1,0.9,length.out=5), using err=0.1. Use plot to visualize the smooth effects corresponding to each quantile. You can set allTerms=TRUE to plot also the parametric effects. How do the smooth and parametrics effects differ between quantiles? NB: here we are plotting the smooth effects, not the predicted quantiles, hence the effects corresponding to, say, quantile 0.9 can fall below that of quantile 0.1.

5. Now we check the median fit. If the output of mqgamV is called fitM then the median fit is fitM[[3]]. Check the bias distribution using check(fitM[[3]]). Recall that we expect that, because we are looking at quantile 0.5, around 50% of the residuals should be negative. Use check1D with the l_gridQCheck1D layer to check that the fraction of negative residuals does not depart too much from 0.5 along any of the covariates.

5 Reaction times for Estonian case-inflected nouns

This question is about modelling reaction time, the main variables are:

- Word: Estonian case-inflected nouns;
- Subject: subjects in lexical decision experiment;
- Trial: trial number in the experiment;

- `LogFrequency`: the log-transformed frequency of the inflected word;
- `WordLength`: the length of the word in letters;
- `Age`: the age of the participant in years;
- `RT`: reaction time;
- `RTinv`: RT transformed by -1000/RT to make it more Gaussian-like
- `InfFamSize`: inflectional family size: the number of different case endings of a noun that are in actual use in the language;

Questions:

1. Load `mgcViz` and the data with `load("data/est.rda")`. Use `qgamV` to fit an additive quantile regression model for the median, with linear effects for `InfFamSize`, `Age`, `LogFrequency`, `WordLength` and `Trial`. Is the effect for `Trial` significant (use `summary`)?
2. Use `check1D` with the `l_gridQCheck1D` layer to check that the fraction of negative residuals does not depart too much from 0.5 along `Trial` and `Subject`. Do you see a pattern in the deviations?
3. Refit the model using a smooth, rather than linear, effect for `Trial` and a random effect for `Subject`. Then check if the effects are significant using `summary` and look at the deviations from 0.5 using `check1D`: are the residual patterns still there?
4. Add a tensor effect (`te(x1,x2)`) for `LogFrequency` and `WordLength`, re-fit and plot all the effects using `plot`. Then get a 3D visualization of the tensor product smooth by extracting the tensor product using the `sm` function, and then plotting it using `plotRGL`. Does this bivariate effect look very non-linear?
5. Substitute the tensor effect with two linear effects, and fit the resulting model to the quantiles `qu = seq(0.1, 0.9, length.out = 5)` using `mqgamV`. Plot the estimated smooth and the random effects using `plot`. Do you see differences in the effect of `Trial` across quantiles? Now using `plot` with `allTerms = TRUE` and `select = 3:6` to plot only the parametric effects (for each quantile). Do you see differences in the estimated effects across different quantiles?

6 Rainfall modelling in Switzerland

This question is about modelling extreme rainfall in Switzerland, mainly using spatio-temporal effects. The main variables are:

- `exra`: the highest rainfall observed in any 12 hour period in that year, in mm;
- `N`: degrees North;
- `E`: degrees East;
- `elevation`: metres above sea level;
- `climate.region`: factor variable indicating one of 12 climate regions;
- `nao`: annual North Atlantic Oscillation index, based on the difference of normalized sea level pressure (SLP) between Lisbon, Portugal and Stykkisholmur/Reykjavik, Iceland. Positive values are generally associated with wetter and milder weather over Western Europe;

- `year`: year of the observation;

Questions:

1. Load `mgcViz`, `gamair` and the data with `data(swer)`. Use `qgamV` to fit an additive quantile regression model for the median, with smooth effects for `nao`, `elevation` and `year` (use `k=5` for the latter), and an isotropic smooth for `E` and `N` (i.e. `s(E,N)`). Look at the significance of the fitted effects using `summary` and plot them using `plot`.

2. We might be interested in verifing whether the rainfall trend is different depending on the climate region. To assess this, modifify the model formula to include by-factor smooth as follows `s(year, climate.region, bs = "fs", k = 5)`. Refit and use `summary` to verify whether the by-region trend term is significant, and plot the by-region trends by extracting it using `sm` and the `l_fitLine(alpha = 1)` layer.

3. We can also verify whether the bivariate spatial effect changes with time, by creating a tensor product between the 2D effect of `E` and `N`, and the effect of `year`. Such an effect can be set up using `te(E, N, year, d = c(2, 1), k = c(20, 5))`. Fit the corresponding median QGAM model, and plot several slices of the 3D tensor product across `year`, using the `plotSlice` function with the `l_fitRaster` and `l_fitContour` layers.

4. Visualize individual 2D slices (across `year`) of the 3D spatio-temporal smooth using the `plotRGL` function (see `?plotRGL.mgcv.smooth.MD` for examples).

5. Go back to the simpler model formula used in the first question and fit the corresponding model to the quantiles `qu = seq(0.1, 0.9, length.out = 9)`, using `mqgamV`. Plot only the univariate effects using `plot` and its `select` argument, and see how they differ between quantiles. Do the same for the spatial effect and for the effect of the climate region.