universität
innsbruck

Transition Models

for Precipitation Climatology Estimation

Nikolaus Umlauf, Reto Stauffer

http://nikum.org

A Journey in Probabilistic Modeling

- Over the past 15 years, we have developed efficient algorithms and open-source software for (Bayesian) distributional regression, culminating in the CRAN package *bamlss*.
- Distributional regression provides a highly flexible framework for full probabilistic modeling of complex data.
- Applications span a wide range: univariate and multivariate responses, count data, censored and survival outcomes, joint models, and more.
- At UIBK, long-term research projects focus on count data modeling, e.g., for probabilistic forecasting of lightning strike counts.

Model specification

Any parameter of a population distribution $\mathcal{D}$ may be modeled by explanatory variables

$$y \sim \mathcal{D}\left(\theta_1(\mathbf{x}; \boldsymbol{\beta}_1),\ \ldots,\ \theta_K(\mathbf{x}; \boldsymbol{\beta}_K)\right),$$

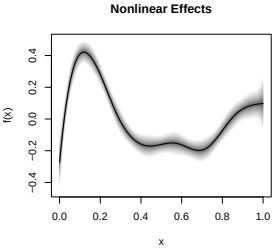
with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top$.

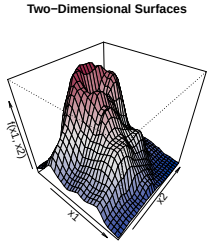
Each parameter is linked to a structured additive predictor

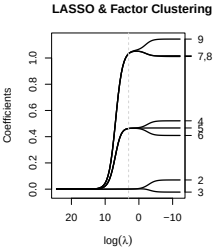
$$h_k(\theta_k(\mathbf{x}; \boldsymbol{\beta}_k)) = f_{1k}(\mathbf{x}; \boldsymbol{\beta}_{1k}) + \ldots + f_{J_k k}(\mathbf{x}; \boldsymbol{\beta}_{J_k k});\ \ j = 1, \ldots, J_k;\ \ k = 1, \ldots, K.$$

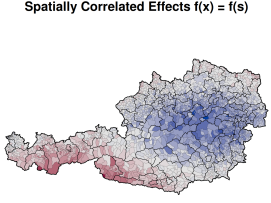
- $h_k(\cdot)$: Link functions for each distribution parameter.
- $f_{jk}(\cdot)$: Model terms of one or more variables.

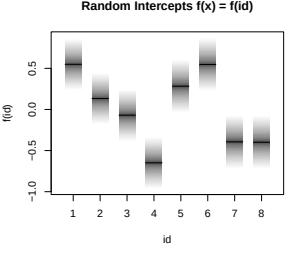
Model Terms $f_{jk}(\cdot)$

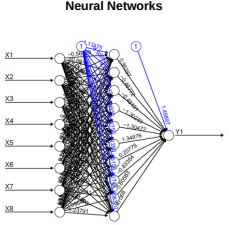






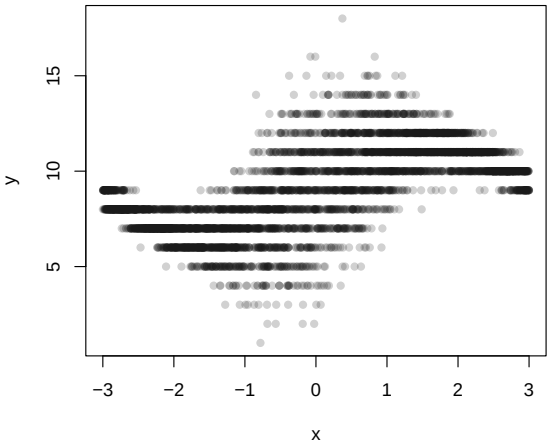






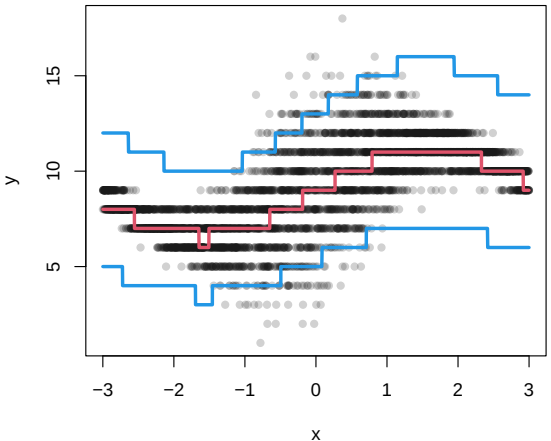
Count Models

Simulated example: $y \sim \text{NO}(\mu = f(x), \log(\sigma) = f(x))$, transformed to counts.



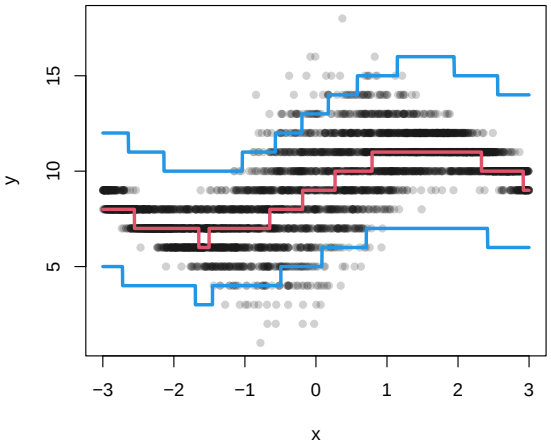
Count Models

Estimated model: $y \sim \text{PO}(\log(\mu) = f(x))$.



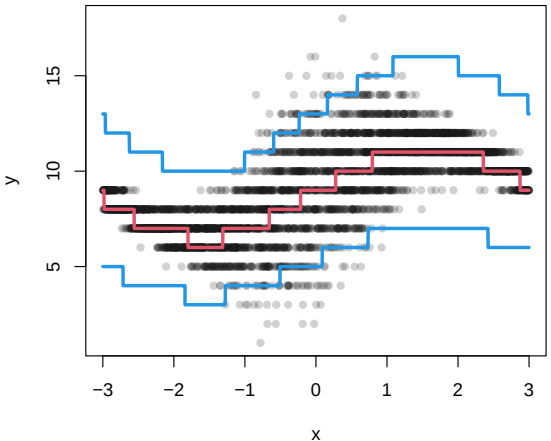
Count Models

Estimated model: $y \sim \text{NBI}(\log(\mu) = f(x), \log(\sigma) = f(x))$.



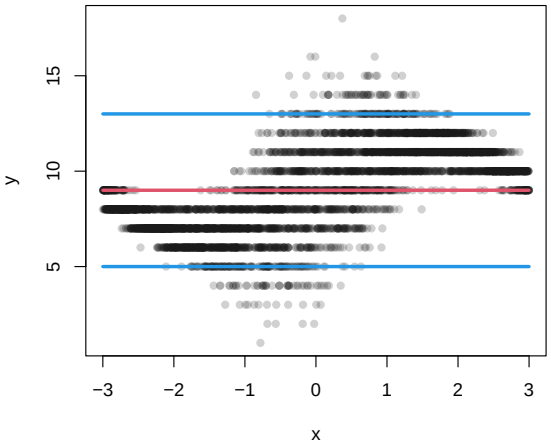
Count Models

Estimated model: $y \sim \text{SICHEL}(\log(\mu) = f(x), \log(\sigma) = f(x), \nu = f(x))$.



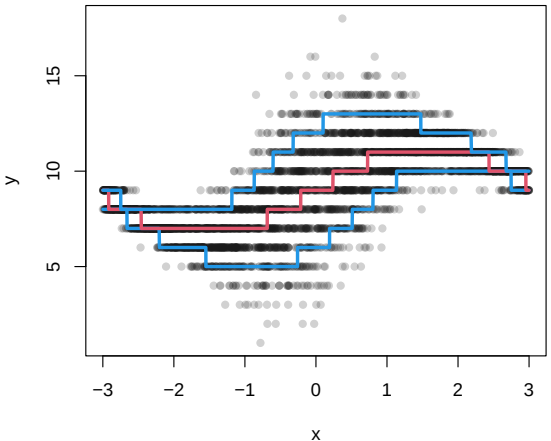
Count Models

Estimated model: $y \sim \text{GPO}(\log(\mu) = f(x), \log(\sigma) = f(x))$.



Count Models

Estimated model: $y \sim \text{DPO}(\log(\mu) = f(x), \log(\sigma) = f(x))$.



Transition Models

The transition probability $P(\cdot)$ for count data is defined as

$$P(y_i > r | y_i \geq r, \mathbf{x}_i) = F(\eta_{ir}(\boldsymbol{\alpha})), \quad r = 0, 1, 2, \ldots$$

where $F(\cdot)$ is a CDF (e.g., logistic or probit) and r represents the counts, with an additive predictor

$$\eta_{ir}(\boldsymbol{\alpha}) = \theta_r + \sum_{j=1}^{k} f_j(\mathbf{x}_i, r; \boldsymbol{\beta}).$$

The parameters $\boldsymbol{\alpha} = (\theta^\top, \beta^\top)$ include count-specific intercepts and (possibly) smooth functions $f_j(\cdot)$. For i.i.d. observations, let π_{ir} denote the probability that the count response equals r, i.e., $P(y_i = r | \mathbf{x}_i)$. These probabilities are computed recursively as

$$\pi_{ir} = (1 - F(\eta_{ir}(\boldsymbol{\alpha}))) \prod_{s=0}^{r-1} F(\eta_{is}(\boldsymbol{\alpha})).$$

Transition Models

Parameter estimation considers the underlying Markov chain $Y_{i0}, Y_{i1}, \ldots$, where

$$Y_{ir} = 1 - I(y_i = r).$$

Simplifies to binary model log-likelihood

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{s=0}^{y_i} \Big[Y_{is} \log(F(\eta_{ir})) + (1 - Y_{is}) \log(1 - F(\eta_{ir})) \Big].$$

$(Y_{i0}, \ldots, Y_{iy_i})^\top = (1, \ldots, 1, 0)$ are created, along with a new covariate $\theta_i = (0, 1, 2, \ldots, y_i)^\top$ to capture count-specific effects $f_j(\mathbf{x}_i, \theta_i)$, or simple count-specific intercepts.

All other covariates are duplicated accordingly.

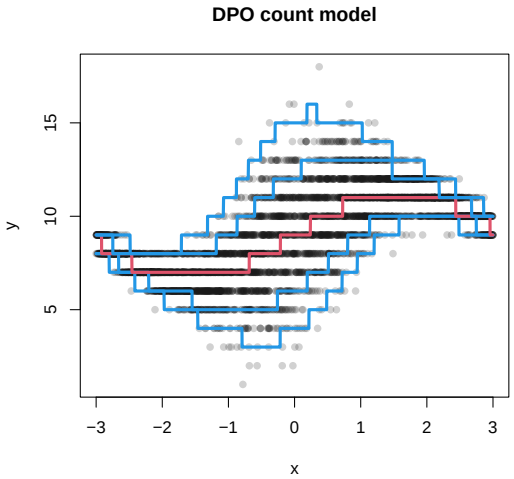
Example

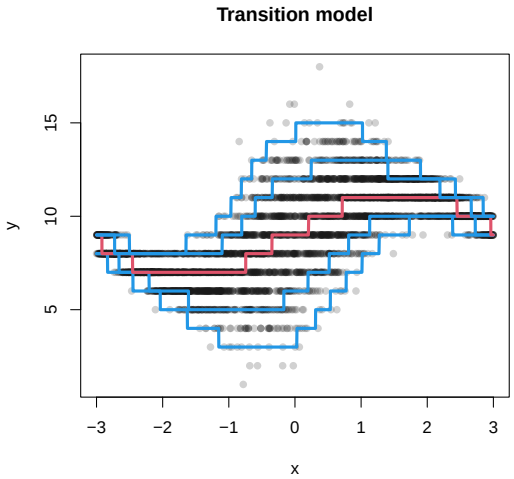
In R:

```
R> print(head(df, 10))
   index myresponse Y theta    x   y   z
1      1          5 1     0 -1.2 765 foo
2      1          5 1     1 -1.2 765 foo
3      1          5 1     2 -1.2 765 foo
4      1          5 1     3 -1.2 765 foo
5      1          5 1     4 -1.2 765 foo
6      1          5 0     5 -1.2 765 foo
7      2          0 0     0  3.2 731 bar
8      3          2 1     0 -0.5 353 foo
9      3          2 1     1 -0.5 353 foo
10     3          2 0     2 -0.5 353 foo

R> b <- glm(Y ~ as.factor(theta) + x + y + z, data = df, family = binomial)
```

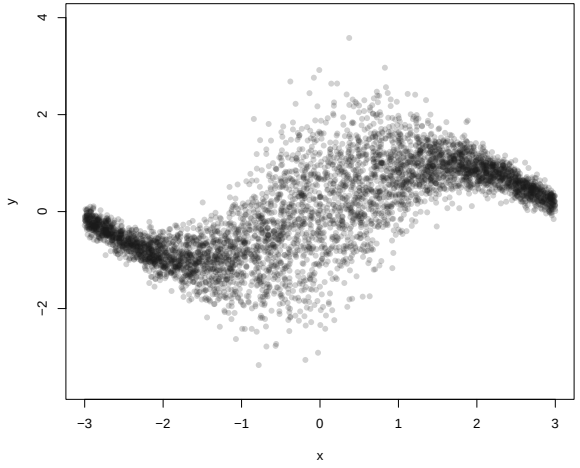
Count Models





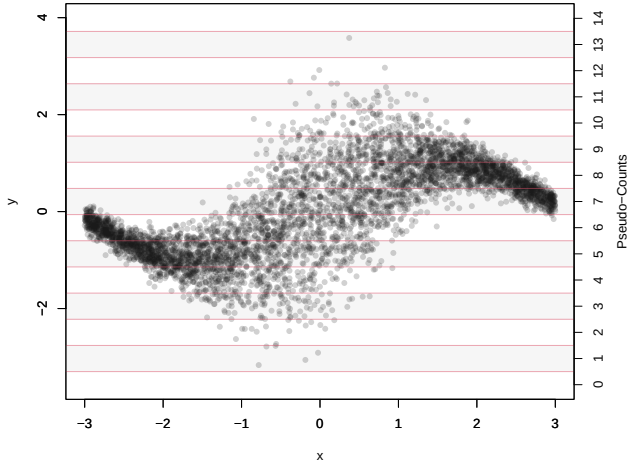
Extension for Continuous Responses

Simulated example: $y \sim \text{NO}(\mu = f(x), \log(\sigma) = f(x))$.



Extension for Continuous Responses

Idea: Define intervals and generate pseudo-counts.



Extension for Continuous Responses

Discretization approach inspired by histogram binning.

Divide response $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, into $m - 1$ intervals using

$$\zeta_1, \zeta_2, \ldots, \zeta_m,$$

where each interval $[\zeta_r, \zeta_{r+1})$ corresponds to a discrete count r.

Each observation y_i is assigned a pseudo count $\tilde{y}_i$.

For a continuous response variable y_i with CDF $F(y)$, discretization process approximates the probabilities of y_i falling into each interval as

$$P(\zeta_r \leq y_i < \zeta_{r+1}) = F(\zeta_{r+1}) - F(\zeta_r).$$

Extension for Continuous Responses

Probabilities are encoded as transformed counts $\tilde{y}_i$, so that the transition model uses

$$P(\tilde{y}_i = r) = P(\zeta_r \leq y_i < \zeta_{r+1})$$

to approximate the discrete likelihood.

The transition model estimates the probability of transitions between counts

$$P(\tilde{y}_i > r \mid \tilde{y}_i \geq r, \mathbf{x}_i) = F(\eta_{ir}(\boldsymbol{\alpha}))$$

and recursively computes

$$P(\tilde{y}_i = r, \mathbf{x}_i) = P(\tilde{y}_i = r \mid \tilde{y}_i \geq r, \mathbf{x}_i) \prod_{s=0}^{r-1} P(\tilde{y}_i > s \mid \tilde{y}_i \geq s, \mathbf{x}_i).$$

Extension for Continuous Responses

- Let r denote the unique index such that $y_i \in [\zeta_r, \zeta_{r+1})$.
- For any value $y_i \in [\zeta_r, \zeta_{r+1})$, the CDF can be approximated by

$$\hat{F}(y_i) = \sum_{s=0}^{r-1} P(\tilde{y}_i = s) + \frac{y_i - \zeta_r}{\zeta_{r+1} - \zeta_r} P(\tilde{y}_i = r).$$

- The PDF can be approximated as

$$\hat{f}(y_i) = \frac{P(\tilde{y}_i = r)}{\zeta_{r+1} - \zeta_r}.$$

- The mean and variance are approximated using midpoints $c_r = \frac{\zeta_r + \zeta_{r+1}}{2}$

$$E[Y] = \sum_r c_r P(\tilde{y} = r), \quad \text{Var}(Y) = \sum_r c_r^2 P(\tilde{y} = r) - \left(\sum_r c_r P(\tilde{y} = r) \right)^2.$$

Extension for Continuous Responses

- Similarly: Skewness $= \frac{E[(Y-E[Y])^3]}{\mathrm{Var}(Y)^{3/2}}$, Kurtosis $= \frac{E[(Y-E[Y])^4]}{\mathrm{Var}(Y)^2}$.
- The τ-quantile $\hat{Q}(\tau)$ is obtained by finding the smallest index r such that

$$\sum_{s=0}^{r} P(\tilde{y} = s) \geq \tau,$$

using linear interpolation within the interval

$$\hat{Q}(\tau) = \zeta_r + \frac{\tau - \sum_{s=0}^{r-1} P(\tilde{y} = s)}{P(\tilde{y} = r)} \cdot (\zeta_{r+1} - \zeta_r).$$

- The mode $\hat{M}$ is approximated as

$$\hat{M} = c_{r^*}, \quad \text{where} \quad r^* = \arg\max_r P(\tilde{y} = r).$$

Software

An implementation is provided in the R package *transitreg*.

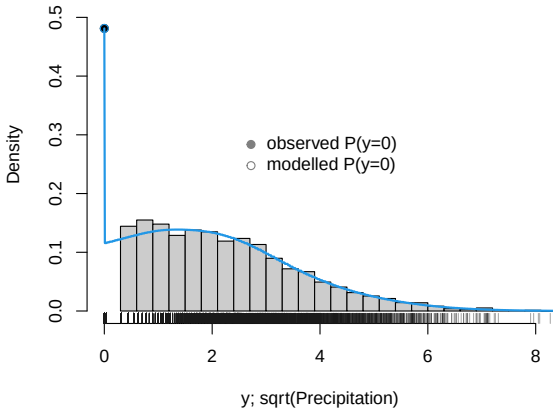
`https://github.com/retostauffer/transitreg`

Install with:

```
R> install.packages("transitreg",
+    repos = c("https://gamlss-dev.R-universe.dev",
+              "https://cloud.R-project.org"))
```

Application: Precipitation Climatology Estimation

```
R> b <- transitreg(sqrt_pre ~ s(theta), data = df, breaks = 30, censored = "left")
```



Application: Precipitation Climatology Estimation

Model formula.

```
R> f <- sqrt_pre ~ theta0 + s(theta, k = 20) + s(day, bs = "cc", k = 20) +
+    te(theta, day, bs = c("cr", "cc"), k = 10)
```

Estimate model.

```
R> breaks <- seq(0, 10, by = 0.05)
R> b <- transitreg(f, data = dtrain, breaks = breaks, censored = "left")
```

Predict quantiles.

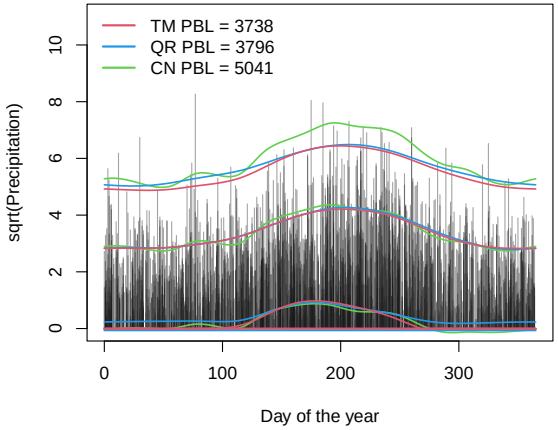
```
R> p <- predict(b, newdata = dtest, prob = c(0.01, 0.1, 0.5, 0.9, 0.99))
```

Compare with *qgam*.

```
R> library("qgam")
R> qu <- c(0.01, 0.1, 0.5, 0.9, 0.99)
R> m <- mqgam(sqrt_pre ~ s(day, k = 20, bs = "cc"), data = dtrain, qu = qu)
```

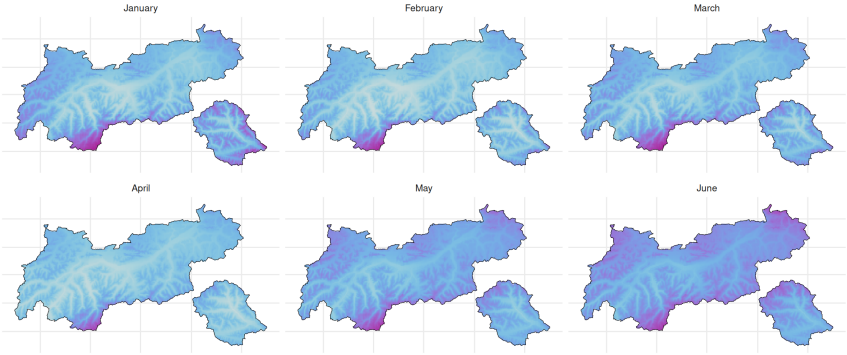
Application: Precipitation Climatology Estimation

Estimated climatology.



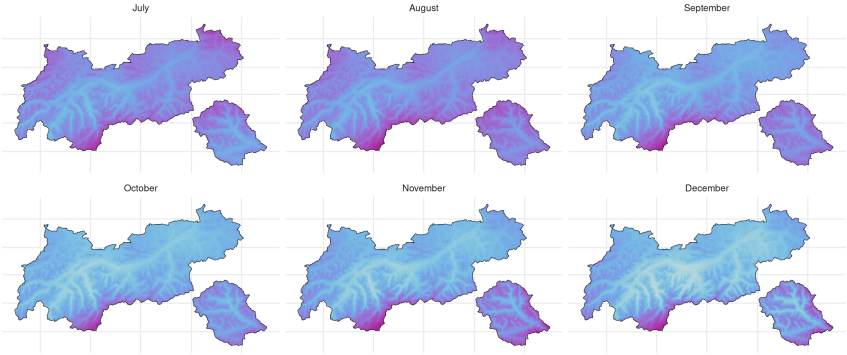
Application: Precipitation Climatology Estimation

Estimated climatology.



Application: Precipitation Climatology Estimation

Estimated climatology.



References

- Berger, M. and Tutz, G. (2021).
  Transition Models for Count Data: A Flexible Alternative to Fixed Distribution Models.
  Statistical Methods and Applications, 30, 1259–1283.
  doi:10.1007/s10260-021-00558-6
- Stauffer, R., Mayr, G.J., Messner, J.W., Umlauf, N., and Zeileis, A. (2017).
  Spatio-Temporal Precipitation Climatology Over Complex Terrain Using a Censored Additive Regression Model.
  International Journal of Climatology, 37(7), 3264–3275.
  doi:10.1002/joc.4913
- Wood, S.N., Goude, Y., and Shaw, S. (2014).
  Generalized Additive Models for Large Data Sets.
  Journal of the Royal Statistical Society: Series C, 64(1), 139–155.
  doi:10.1111/rssc.12068
- Wood, S.N., Li, Z., Shaddick, G., and Augustin, N.H. (2017).
  Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data.
  Journal of the American Statistical Association, 112(519), 1199–1210.
  doi:10.1080/01621459.2016.1195744