technische universität
dortmund

New Methods for Statistical Learning

Andreas Groll

June 16^{th} 2020

Dortmund Data Science Center
Department of Statistics

Short introduction of my work group

- Andreas Groll
- since September 2019: Professor for *Statistical Methods for Big Data* at TU Dortmund University

Short introduction of my work group

- Andreas Groll
- since September 2019: Professor for *Statistical Methods for Big Data* at TU Dortmund University
- Work group: 4 Ph.D. students (one external)

Short introduction of my work group

- Andreas Groll
- since September 2019: Professor for *Statistical Methods for Big Data* at TU Dortmund University
- Work group: 4 Ph.D. students (one external)

Before:

- Assistant Professor for Data Analysis & Statistical Algorithms at TU Dortmund (11/17-08/19)

Short introduction of my work group

- Andreas Groll
- since September 2019: Professor for *Statistical Methods for Big Data* at TU Dortmund University
- Work group: 4 Ph.D. students (one external)

Before:

- Assistant Professor for Data Analysis & Statistical Algorithms at TU Dortmund (11/17-08/19)
- Postdoc at Georg-August University Göttingen (03/16-10/17) and LMU Munich (04/12-02/16)

Short introduction of my work group

- Andreas Groll
- since September 2019: Professor for *Statistical Methods for Big Data* at TU Dortmund University
- Work group: 4 Ph.D. students (one external)

Before:

- Assistant Professor for Data Analysis & Statistical Algorithms at TU Dortmund (11/17-08/19)
- Postdoc at Georg-August University Göttingen (03/16-10/17) and LMU Munich (04/12-02/16)
- 4 research stays at Stanford University (01-04/15; 08/17; 02-03/18; 09/19)

Short introduction of my work group

- Andreas Groll
- since September 2019: Professor for *Statistical Methods for Big Data* at TU Dortmund University
- Work group: 4 Ph.D. students (one external)

Before:

- Assistant Professor for Data Analysis & Statistical Algorithms at TU Dortmund (11/17-08/19)
- Postdoc at Georg-August University Göttingen (03/16-10/17) and LMU Munich (04/12-02/16)
- 4 research stays at Stanford University (01-04/15; 08/17; 02-03/18; 09/19)
- Dissertation (2011) in Statistics (LMU Munich)

Short introduction of my work group

- Andreas Groll
- since September 2019: Professor for *Statistical Methods for Big Data* at TU Dortmund University
- Work group: 4 Ph.D. students (one external)

Before:

- Assistant Professor for Data Analysis & Statistical Algorithms at TU Dortmund (11/17-08/19)
- Postdoc at Georg-August University Göttingen (03/16-10/17) and LMU Munich (04/12-02/16)
- 4 research stays at Stanford University (01-04/15; 08/17; 02-03/18; 09/19)
- Dissertation (2011) in Statistics (LMU Munich)
- Diploma (2007) in Business Mathematics (LMU Munich)

Research interests (general)

- Methods for variable selection and regularization, in particular in Generalized Linear/Additive (Mixed) Models and time-to-event data analysis
- Modeling of categorical data
- Semiparametric regression
- Sports Statistics, in particular modeling and prediction of international football tournaments

Effects selection in Cox frailty models

Cox frailty model with time-varying coefficients:

$$\lambda(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \lambda_0(t) \exp\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \sum_{k=1}^{r} z_{ijk}\gamma_k(t) + \boldsymbol{u}_{ij}^T\boldsymbol{b}_i,\right)$$

with covariates $z_{ij1}, \dots, z_{ijr}$ being associated with time-varying effects.

Effects selection in Cox frailty models

Cox frailty model with time-varying coefficients:

$$\lambda(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \lambda_0(t) \exp\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \sum_{k=1}^{r} z_{ijk}\gamma_k(t) + \boldsymbol{u}_{ij}^T\boldsymbol{b}_i,\right)$$

with covariates $z_{ij1}, \ldots, z_{ijr}$ being associated with time-varying effects.

Estimation: expand time-varying effects $\gamma_k(t)$ in B-splines:

$$\gamma_k(t) = \sum_{m=1}^{M} \alpha_{k,m} B_m(t; d)$$

Effects selection in Cox frailty models

Cox frailty model with time-varying coefficients:

$$\lambda(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \lambda_0(t) \exp\left(\boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \sum_{k=1}^{r} z_{ijk}\gamma_k(t) + \boldsymbol{u}_{ij}^T \boldsymbol{b}_i, \right)$$

with covariates $z_{ij1}, \ldots, z_{ijr}$ being associated with time-varying effects.

Estimation: expand time-varying effects $\gamma_k(t)$ in B-splines:

$$\gamma_k(t) = \sum_{m=1}^{M} \alpha_{k,m} B_m(t; d)$$

- Classical variable selection via LASSO
- Effects selection of potential time-varying coefficients via L_1-penalization
- Effects selection of potential time-varying coefficients via boosting

Interpretable Machine Learning

⟹ Learn behavior of model by observing changes in output, while changing input, e.g. finding classes modeled next to each other

Interpretable Machine Learning

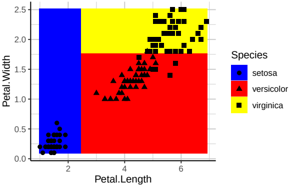
⟹ Learn behavior of model by observing changes in output, while changing input, e.g. finding classes modeled next to each other

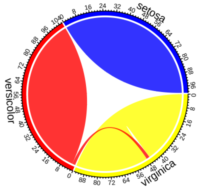
1. A good (ML) model is fitted on `iris` data
2. Labels are predicted for all observations
3. The value of the feature *Petal.Width* is raised by a very small amount for all observations
4. New labels are predicted for the manipulated data
5. Changes found are interpreted as "classes modeled next to each other" and visualized

Interpretable Machine Learning

$\Longrightarrow$ Learn behavior of model by observing changes in output, while changing input, e.g. finding classes modeled next to each other

1. A good (ML) model is fitted on `iris` data
2. Labels are predicted for all observations
3. The value of the feature *Petal.Width* is raised by a very small amount for all observations
4. New labels are predicted for the manipulated data
5. Changes found are interpreted as "classes modeled next to each other" and visualized





Penalized Joint Regression Modeling

Bivariate count obervations (y_1, y_2) and marginal regressions:

- $\hat{y}_1 = \exp(\beta_0^{(1)} + \beta_1^{(1)} x_1^{(1)} + \ldots + \beta_p^{(1)} x_p^{(1)})$
- $\hat{y}_2 = \exp(\beta_0^{(2)} + \beta_1^{(2)} x_1^{(2)} + \ldots + \beta_p^{(2)} x_p^{(2)})$

With dependency between y_1 and y_2 taken into account via Copulae C

- $F(\hat{y}_1, \hat{y}_2) = C(F_1(\hat{y}_1), F_2(\hat{y}_2))$

Penalized Joint Regression Modeling

Bivariate count obervations (y_1, y_2) and marginal regressions:

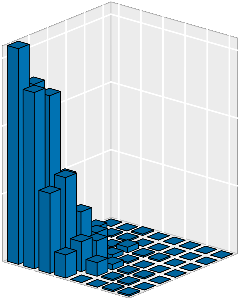
- $\hat{y}_1 = \exp(\beta_0^{(1)} + \beta_1^{(1)} x_1^{(1)} + \ldots + \beta_p^{(1)} x_p^{(1)})$
- $\hat{y}_2 = \exp(\beta_0^{(2)} + \beta_1^{(2)} x_1^{(2)} + \ldots + \beta_p^{(2)} x_p^{(2)})$

With dependency between y_1 and y_2 taken into account via Copulae C

- $F(\hat{y}_1, \hat{y}_2) = C(F_1(\hat{y}_1), F_2(\hat{y}_2))$

New: Penalization for competitive settings

- $\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\xi \sum_{j=0}^{p} \omega_j \left(\beta_j^{(1)} - \beta_j^{(2)}\right)^2$



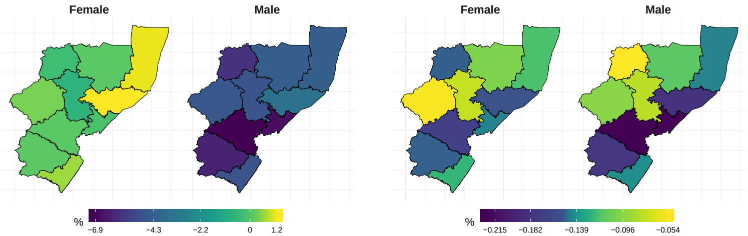
$\Longrightarrow$ Probabilities for football scores (y_1, y_2) and interpretable coefficients

Regularization in complex regression settings

1. Causal inference using Distributional Regression with instrumental variables.
   - Causal effects of treatments on various distributional quantities, e.g. expectation, variance, coefficient of variation.
   - Data-driven variable selection via gradient-based-boosting.

Regularization in complex regression settings

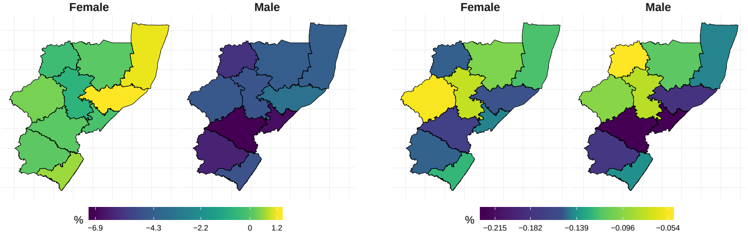
1. Causal inference using Distributional Regression with instrumental variables.
   - Causal effects of treatments on various distributional quantities, e.g. expectation, variance, coefficient of variation.
   - Data-driven variable selection via gradient-based-boosting.



Effect of electricity on mean (left) and standard deviation (right) on employment rates.

Regularization in complex regression settings

1. Causal inference using Distributional Regression with instrumental variables.
   - Causal effects of treatments on various distributional quantities, e.g. expectation, variance, coefficient of variation.
   - Data-driven variable selection via gradient-based-boosting.

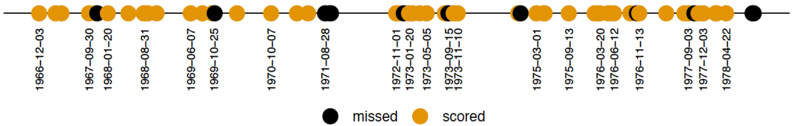


Effect of electricity on mean (left) and standard deviation (right) on employment rates.

2. Regularized bivariate mixed binary-continuous copula regression.
   - Joint modeling of e.g. match winner and match duration in tennis.

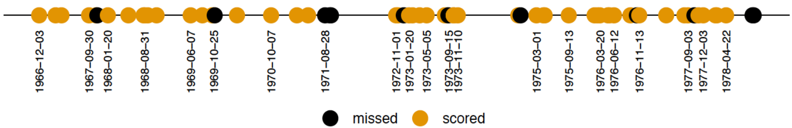
Modeling the “hot hand” effect in sports via HMMs

Modeling the “hot hand” effect in sports via HMMs

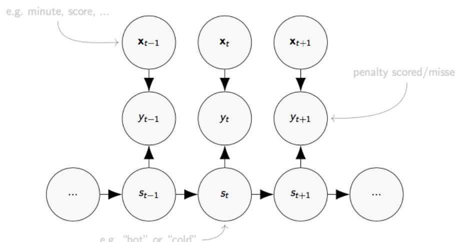


Example data – penalties taken by Gerd Müller

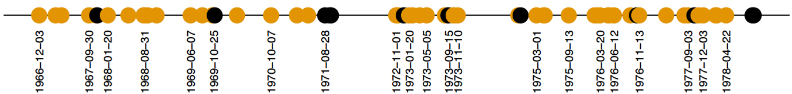
Modeling the “hot hand” effect in sports via HMMs



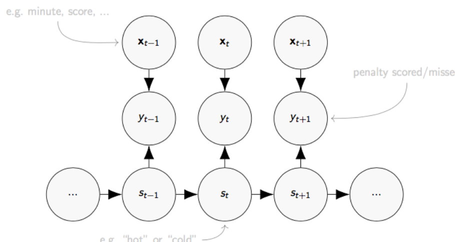
Example data – penalties taken by Gerd Müller



Modeling the “hot hand” effect in sports via HMMs



Example data – penalties taken by Gerd Müller



- Planned: Analysis on “hot glove” effect

Current grant proposals

- Gemeinsamer Bundesausschuss Innovationsausschuss: PREMISE (ongoing)
- Marie-Curie ITN: S-TRAINING (recently rejected; currently under revision)
- 2 DFG Research Training Groups:
  - Biostatistical methods for high-dimensional data in toxicology (submitted)
  - Domain knowledge in the data-driven sciences from basic research to industrial applications (submitted)

Planned grant proposals

- 2 DFG Research Units
- DFG Individual Research Grant
- DFG Collaborative Research Centre

References

Briseño Sánchez, G., Hohberg, M., Groll, A. & Kneib, T. (2019): Flexible Instrumental Variable Distributional Regression, Proceedings of the 34th International Workshop on Statistical Modelling, Volume **2**, 299–304.

Briseño Sánchez, G. & Groll, A. (2020). Modelling the effect of rural electrification on employment via component-wise boosted causal distributional regression. Proceedings of the 35th International Workshop on Statistical Modelling, to appear.

Groll, A., Hastie, T. & Tutz, G. (2017). Selection of effects in cox frailty models by regularization methods, *Biometrics* **73**(3), 846–856.

Groll, A., Hastie, T., Kneib, T. & Tutz, G. (2018): Boosting Methods for Effects Selection in Cox Frailty Models, Proceedings of the 33rd International Workshop on Statistical Modelling, Volume **1**, 122–127.

References

- Groll, A. & Hohberg, M. (2019): An adaptive lasso Cox frailty model for time-varying covariates based on the full likelihood, Proceedings of the 34th International Workshop on Statistical Modelling, Volume 1, 67–72.

- Ötting, M. & Groll, A. (2019): A regularized hidden Markov model for analyzing the 'hot shoe' in football, *arXiv preprint* arXiv:1911.08138.

- Ötting, M. & Groll, A. (2020): Regularisation in hidden Markov models with an application to football data, Proceedings of the 35th International Workshop on Statistical Modelling, to appear.

- Van der Wurp, H., Groll, A., Kneib, T., Marra, G. & Radice, R. (2020): Generalised Joint Regression for Count Data - A Penalty Extension for Competitive Settings. *Statistics and Computing*, to appear.