HelmholtzZentrum münchen
German Research Center for Environmental Health

Bayesian Nonparametric Clustering and Inference for Inpatient Health Care Utilization

Christoph Kurz, Laura Hatfield

Helmholtz Zentrum München
Harvard Medical School

HELMHOLTZ
ASSOCIATION

Background

Inpatient hospital services account for a **small share** of health care utilization but the **majority** of total health care spending.

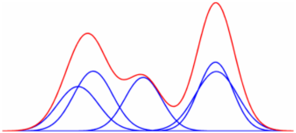
- What are the driving forces of inpatient health care spending? (**inference**, **interpretation**)
- How can we account for different patient characteristics (**subgroup analysis**, **clustering**)

Background

Mixture distributions are good way to model health care utilization

A mixture distribution f_{mix} is a weighted sum, $\Sigma c_i = 1$, of a finite set of probability density functions $p_1(x)$, ..., $p_k(x)$

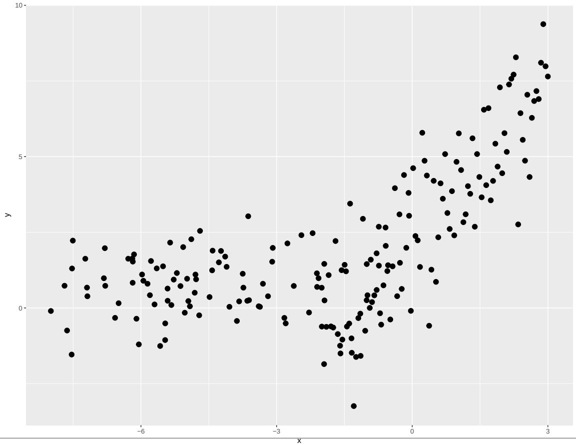
$$f_{mix}(x) = \sum_{i=1}^{K} c_i \, p_i(x).$$



They can account for zero-inflation, over-dispersion, and skewness.

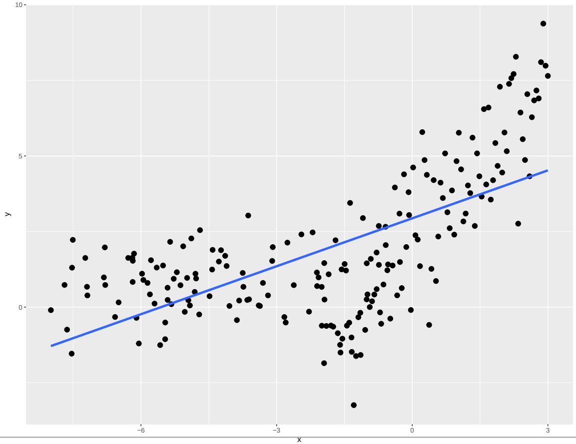
Background

Mixture models can be extended to regression



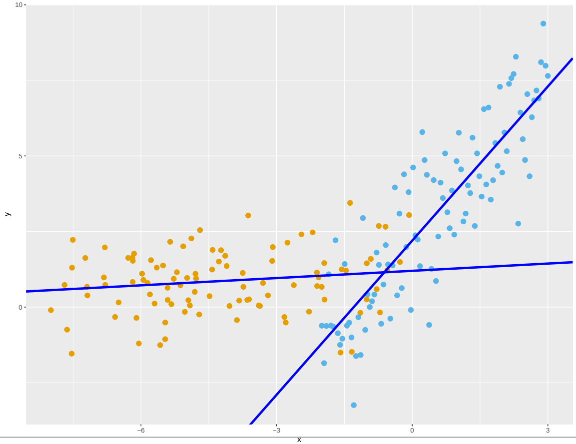
Background

Mixture models can be extended to regression



Background

Mixture models can be extended to regression



Background

The are two ways to specify the number of mixture components (= clusters)

- Specify the number of components before the analysis (*ex-ante*).
- Calculate different models with different clusters and select the "best" (*ex-post*).

Both methods introduce a **decision-bias** and **modelselection-bias**.

Methods

Bayesian nonparametric models allow to estimate the number of components K from the data.

- avoids over- and underfitting
- model only as complex as the data require
- in theory, model complexity is unbounded (infinite number of clusters)

Methods

We developed a Dirichlet Process mixture regression model for counts (hospital days), DP-NB

$$\underbrace{y|X}_{days} \sim \sum_{k=1}^{K} \underbrace{c_k|X}_{weights} \cdot \underbrace{\text{NegBin}(\mu_k, \psi_k)}_{regression\ model},$$

with

$$\mu_k = \exp(X\beta_k).$$

We also extend this model to a zero-inflated version (DP-ZINB).

Simulation Study

The DP-NB finds the true number of components more accurately than AIC and BIC selection methods

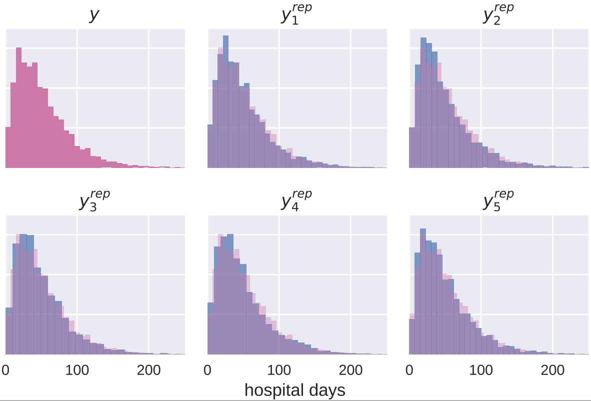
	high overlap			medium overlap			low overlap		
Truth	AIC	BIC	**DP-NB**	AIC	BIC	**DP-NB**	AIC	BIC	**DP-NB**
2	5	1	4	3	3	**2**	1	1	3
3	1	1	4	4	4	4	1	1	4
4	1	1	**4**	1	1	3	1	1	5
5	1	1	3	5	1	6	1	1	6

AOK data set

- AOK claims data set with incident lung cancer in 2009 (Schwarzkopf et al., 2015)
- AOK is the largest health insurance company in Germany and covers around a third of the German population
- outcome: **total number of inpatient hospital days** (1 year period)
- only patients who survived the full year where included (N=7118)

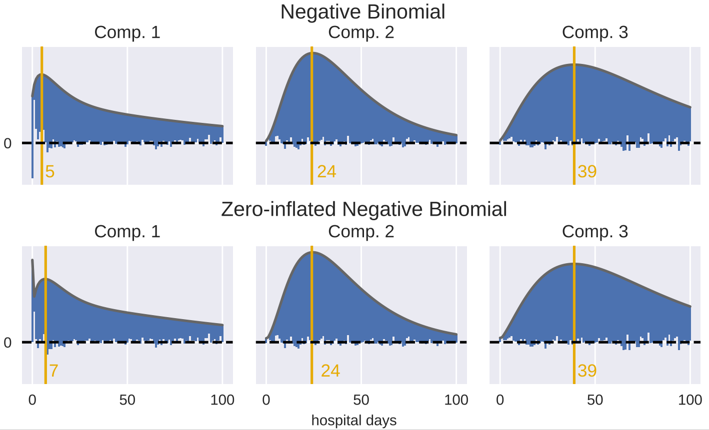
Results

The posterior predictive distribution of replicated outcome y^{rep} is close to the true outcome



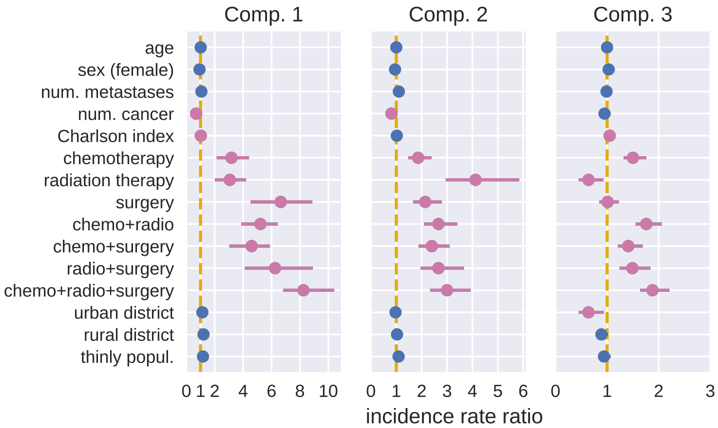
Results

The DP-NB finds three components for the AOK data set



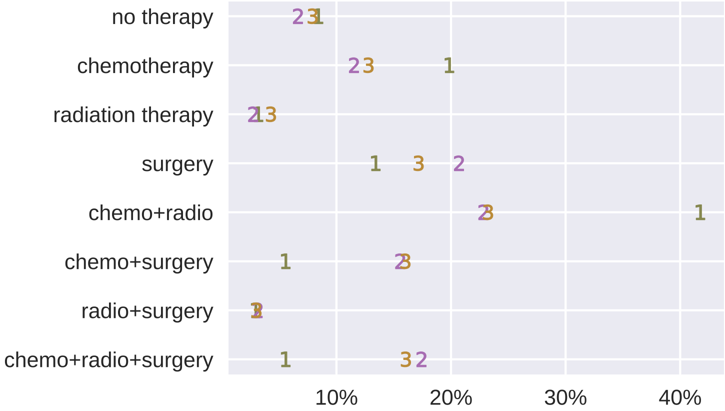
Results

Biggest differences are in treatment coefficients



Results

Component 1 gets the most chemotherapy and the least surgery



Discussion

Component 1 has patients in more advanced stages of lung cancer

- less hospital days $\neq$ healthy
- less surgery, but more chemotherapy and radiation therapy

Discussion

Component 2 and 3 have more cases with good prospect

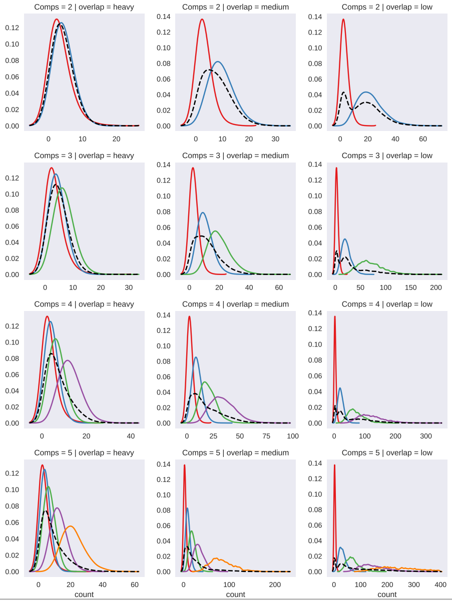
- more surgery
- more surgery + chemotherapy + radiation therapy
- Component 3 is very similar to Component 2 but has individuals with more comorbidities and who are older.

Conclusion

- the presented Bayesian clustering and inference method for count data can be used to find subgroups of patients while still being fully interpretable
- because of its non-parametric nature it avoids over- and underfitting of the cluster components.
- on the AOK data set, it can find subgroups with specific properties that correspond well to the different number of hospital days in each component

Thank you

Simulation



Comps = 2 | overlap = heavy
Comps = 2 | overlap = medium
Comps = 2 | overlap = low
Comps = 3 | overlap = heavy
Comps = 3 | overlap = medium
Comps = 3 | overlap = low
Comps = 4 | overlap = heavy
Comps = 4 | overlap = medium
Comps = 4 | overlap = low
Comps = 5 | overlap = heavy
Comps = 5 | overlap = medium
Comps = 5 | overlap = low
count
count
count