Latent Space Network Models

Florian Pargent

28.06.2017

Grundidee des Latent Space Ansatzes

Typische Netzwerk Eigenschaften

- Transitivity
- Homophily
- Community Structure
- Degree Heterogeneity

Modellierung eines latenten “Social Space”

- Jeder Knoten im Netzwerk wird repräsentiert durch einen Positionsvektor im $\mathbb{R}^2$ (selten $\mathbb{R}^3$ oder höher)
- Wahrscheinlichkeit für das Vorliegen einer Kante zwischen zwei Knoten hängt ab vom euklidischen Abstand der beiden Knoten
- Bedingt auf die latenten Positionen der Knoten, sind alle Kanten voneinander unabhängig

Vorteile:

- Modellierbarkeit typischer Charakteristiken von Netzwerken
- Natürliche grafische Darstellung des Netzwerkmodells

Beispieldatensatz: Publikationen am Statistikinstitut

Webscraping von der Homepage des Statistikinstituts

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR MATHEMATIK, INFORMATIK UND STATISTIK
INSTITUT FÜR STATISTIK

Google Benutzerdefinierte | www.lmu.de | Fakultät für Mathematik, Informatik und Statistik | LMU-Portal | Sitemap

Startseite ‣ Forschung ‣ Veröffentlichungen

- AKTUELLES
- INSTITUT
- PERSONEN
- FORSCHUNG
  - Projekte
  - Veröffentlichungen
  - Technical Reports
- STUDIENINTERESSIERTE
- STUDIUM
- PROMOTIONSPROGRAMM
- STELLEN
- KONTAKT

drucken

Veröffentlichungen

Springe zu: 2017 | 2016 | 2015 | 2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2007 | 2005 | 2004 | 2003 | 2001

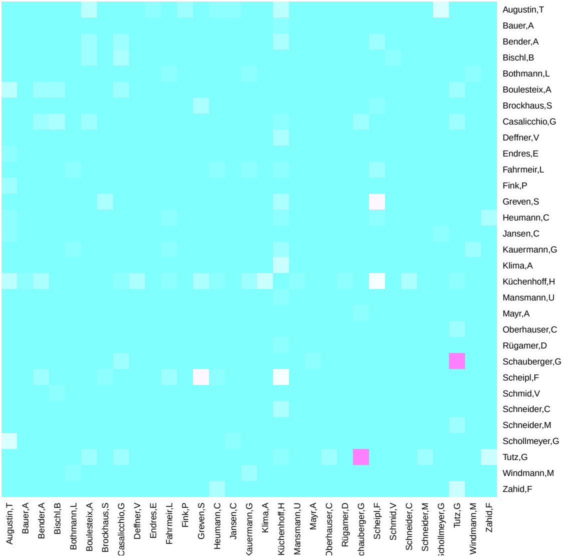
2017

- Brockhaus, Sarah; Melcher, Michael; Leisch, Friedrich; Greven, Sonja (2017)
  Boosting Flexible Functional Regression Models with a High Number of Functional Historical Effects
  In: Statistics and Computing, Vol. 27, Nr. 4: S. 913-926
- Janitza, Silke (2017)
  On the overestimation of random forest's out-of-bag error
  (Volltext vorhanden)

QUICK LINKS

- LSF SoSe 2017
- Frauenbeauftragte
- Statistisches Beratungslabor
- R-Kurse
- Center for Quantitative Risk Analysis
- Blog stats@lmu
- MSc Data Science
- Bibliothek Wirtschaftswissenschaften & Statistik
- Ausgewählte Abschlussarbeiten
- Institutskolloquium

Soziomatrix (31 Knoten)

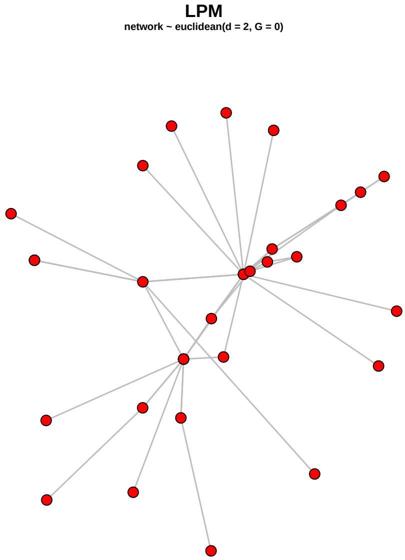


Latent Space Modelle in aufsteigender Komplexität (Anwendung mit dem R Paket latentnet)

Latent Position Model (LPM)

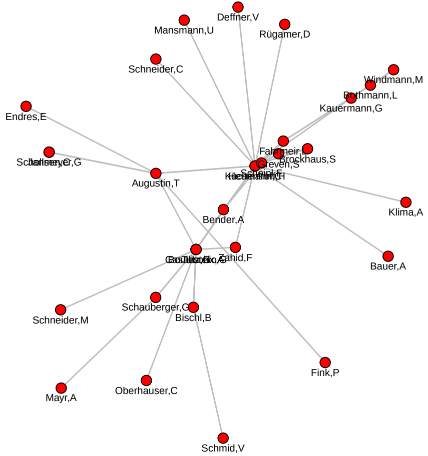
$$log\left(\frac{P(Y_{ij}=1|Z,x,\beta)}{P(Y_{ij}=0|Z,x,\beta)}\right)=\sum_{k=0}^{p}\beta_k x_{kij}-\|Z_i-Z_j\|$$

(Hoff, Raftery, and Handcock 2002)


LPM
network ~ euclidean(d = 2, G = 0)

LPM

network ~ euclidean(d = 2, G = 0)



Latent Cluster Model (LCM)

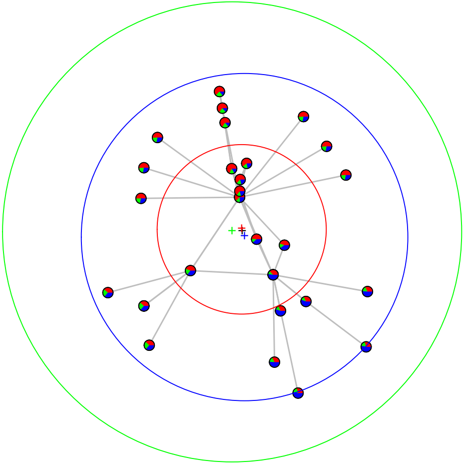
$$log\left(\frac{P(Y_{ij}=1|Z,x,\beta)}{P(Y_{ij}=0|Z,x,\beta)}\right)=\sum_{k=0}^{p}\beta_k x_{kij}-\|Z_i-Z_j\|$$

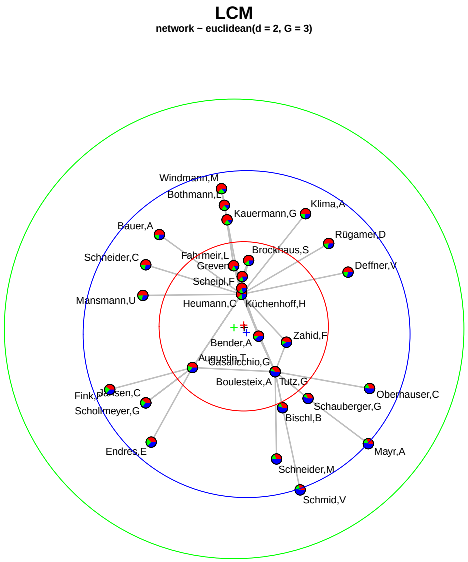
$$Z_i\sim\sum_{g=1}^{G}\lambda_g MVN_d(\mu_g,\,\sigma_g^2 I_d)$$

(Handcock, Raftery, and Tantrum 2007)

LCM

network ~ euclidean(d = 2, G = 3)




LCM
network ~ euclidean(d = 2, G = 3)
Windmann,M
Bothmann,L
Kauermann,G
Klima,A
Bauer,A
Rügamer,D
Brockhaus,S
Schneider,C
Fahrmeir,L
Greven,S
Deffner,V
Scheipl,F
Mansmann,U
Heumann,C
Küchenhoff,H
Zahid,F
Bender,A
Augustin,T
Casalicchio,G
Boulesteix,A
Tutz,G
Oberhauser,C
Fink,P
Jansen,C
Schauberger,G
Schollmeyer,G
Bischl,B
Endres,E
Mayr,A
Schneider,M
Schmid,V

Latent Cluster Random Effects Model (LCREM)

$$log\left(\frac{P(Y_{ij}=1|Z,x,\beta)}{P(Y_{ij}=0|Z,x,\beta)}\right)=\sum_{k=0}^{p}\beta_k x_{kij}-\|Z_i-Z_j\|+\delta_i+\delta_j$$

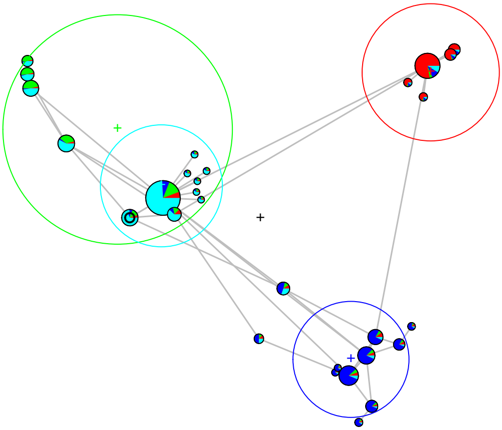
$$\delta_{i,j}\sim N(0,\,\sigma_\delta^2)$$

$$Z_i\sim\sum_{g=1}^{G}\lambda_g MVN_d(\mu_g,\,\sigma_g^2 I_d)$$

(Krivitsky et al. 2009)

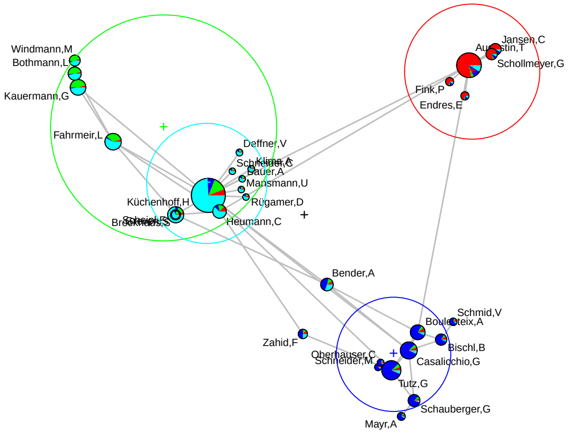
LCREM

network ~ euclidean(d = 2, G = 4) + rsociality



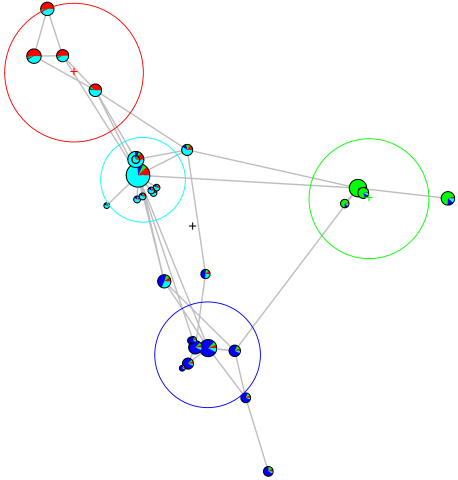
LCREM

network ~ euclidean(d = 2, G = 4) + rsociality



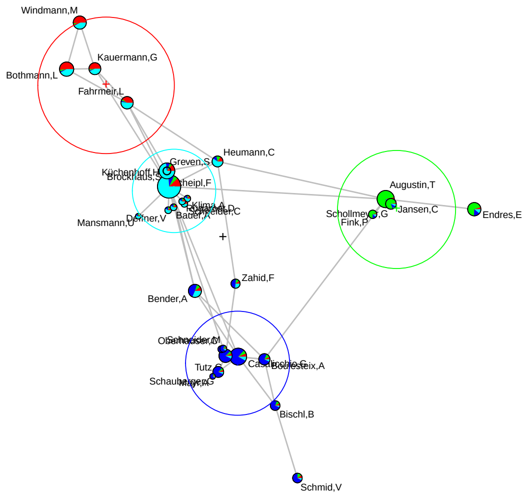
LCREM mit Kovariable Professoren

network ~ euclidean(d = 2, G = 4) + rsociality + edgecov(zero_profs) + edgecov(two_profs)



LCREM mit Kovariable Professoren

network ~ euclidean(d = 2, G = 4) + rsociality + edgecov(zero_profs) + edgecov(two_profs)



Schätzwerte für β

	Posterior Mean	2.5%	97.5%
Intercept (Ein Professor)	0.37	-0.32	1.12
Kein Professor	-1.81	-3.42	-0.32
Zwei Professoren	1.40	-0.06	2.93

Erweiterung auf nicht binäre Netzwerke

Zähldaten

$$Y_{ij}|\mu_{ij} \sim Poisson(\mu_{ij})$$

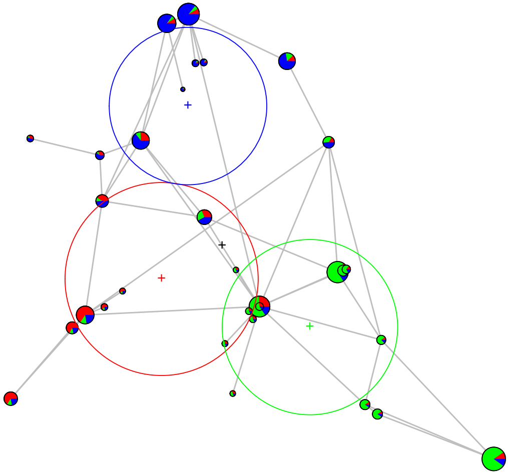
$$log(\mu_{ij}) = \sum_{k=0}^{p} \beta_k x_{kij} - \|Z_i - Z_j\| + \delta_i + \delta_j$$

$$\delta_{i,j} \sim N(0,\, \sigma_\delta^2)$$

$$Z_i \sim \sum_{g=1}^{G} \lambda_g MVN_d(\mu_g,\, \sigma_g^2 I_d)$$

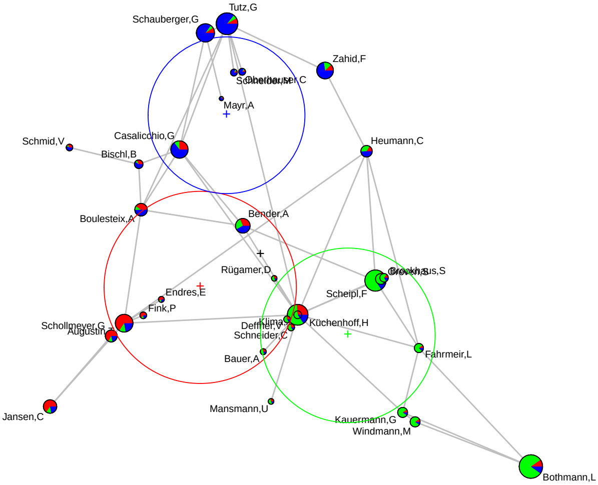
LCREM mit Poisson Response und Kovariable Professoren

network ~ euclidean(d = 2, G = 3) + rsociality + edgecov(zero_profs) + edgecov(two_profs)   (Weight)



LCREM mit Poisson Response und Kovariable Professoren

network ~ euclidean(d = 2, G = 3) + rsociality + edgecov(zero_profs) + edgecov(two_profs)   (Weight)



Schätzwerte für β

	Posterior Mean	2.5%	97.5%
Intercept (Ein Professor)	0.51	-0.05	1.04
Kein Professor	-1.61	-2.82	-0.42
Zwei Professoren	0.87	-0.25	1.99

Modellschätzung

ML Schätzung am Beispiel des LPM

log-Likelihood (vgl. logistische Regression):

$$log(f(y|\eta)) = \sum_{i \neq j} (\eta_{ij} y_{iy} - log(1 + e^{\eta_{ij}}))$$

$$\eta_{ij} = \sum_{k=0}^{p} \beta_k x_{kij} - \|Z_i - Z_j\| = \sum_{k=0}^{p} \beta_k x_{kij} - d_{ij}$$

Likelihood ist konkav bzgl. D aber nicht bzgl. Z.

Deshalb numerische Optimierung:

- Approximiere D durch geodesische Distanzen
- Identifiziere Positionen Z aus D mit multidimensionaler Skalierung
- Verwende Z als Startwerte für nicht lineare Optimierung (z.b. optim in R)

Volle Bayes Inferenz am Beispiel des LCM

Priori Verteilungen

$$\beta \sim MVN_{p+1}(\xi, \Psi)$$

$$\lambda \sim Dirichlet(\nu)$$

$$\mu_g \sim MVN_d(0, \omega^2 I_d)$$

$$\sigma_g^2 \sim \sigma_0^2 Inv\chi_\alpha^2$$

Full Conditionals (bekannt bis auf Konstanten)

$$z_i|K_i, \ldots \propto \phi_d(z_i; \mu_g, \sigma_g^2 I_d)\, f(y|Z, x, \beta)$$

$$\beta|Z, \ldots \propto \phi_{p+1}(\beta; \xi, \Psi)\, f(y|Z, x, \beta)$$

Full Conditionals (bekannt)

$$\lambda|\ldots \sim Dirichlet(m + \nu)$$

$$\mu_g|\ldots \sim MVN_d\left(\frac{m_g \bar{z}_g}{m_g + \sigma_g/\omega^2}, \frac{\sigma_g^2}{m_g + \sigma_g^2/\omega^2} I\right)$$

$$\sigma_g^2|\ldots \sim (\sigma_0^2 + d s_g^2) Inv\chi^2_{\alpha + m_g d}$$

$$P(K_i = g|\ldots) = \frac{\lambda_g \phi_d(z_i; \mu_g, \sigma_g^2 I_d)}{\sum_{r=1}^{G} \lambda_r \phi_d(z_i; \mu_r, \sigma_r^2 I_d)}$$

MCMC Algorithmus

Metropolis-Algorithmus für Z_{t+1} (zufällige Reihenfolge)

- Vorschlag: $Z_i^* \sim MVN_d(Z_{it}, \tau_Z^2 I_d)$
- Akzeptanz WK: $\frac{f(y|Z^*,x,\beta_t)\phi_d(Z_i^*;\mu_{K_i},\sigma_{K_i}^2 I_d)}{f(y|Z_t,x,\beta_t)\phi_d(Z_{it};\mu_{K_i},\sigma_{K_i}^2 I_d)}$

Metropolis-Algorithmus für β_{t+1}

- Vorschlag: $\beta^* \sim MVN_{p+1}(\beta_t, \tau_\beta^2 I_{p+1})$
- Akzeptanz WK: $\frac{f(y|Z_{t+1},x,\beta^*)\phi_{p+1}(\beta^*;\xi,\Psi)}{f(y|Z_{t+1},x,\beta_t)\phi_{p+1}(\beta_t;\xi,\Psi)}$

Gibbs-Sampling für:

K_i, μ_g, σ_g^2 und λ_g

Identifizierbarkeit der Modellparameter

Latente Positionen

- Die Likelihood hängt von den latenten Positionen nur über deren Distanzen ab
- Damit ist die Likelihood invariant gegenüber Spiegelung, Rotation und Verschiebung der latenten Positionen

Lösung:

Suche die Parameterschätzungen mit dem geringsten erwarteten a posteriori Verlust (bzgl. der Kullback-Leibler Verlustfunktion)

$$\hat{\eta}_{MKL} = \underset{\eta_*}{\operatorname{argmin}} \; E_{\eta|Y_{obs}}(KL(\eta, \eta_*))$$

$$= \underset{\eta_*}{\operatorname{argmax}} \; \frac{\exp(\eta_*^T E_\eta(Y|Y_{obs}))}{\prod_{i \neq j}(1 + \exp(\eta_{ij*}))}$$

(Shortreed, Handcock, and Hoff 2006):

Klassen Labels

- Invarianz gegenüber Permutationen der Clusterlabels
- Labelswitching Problem analog zu anderen Mixture Modellen

Lösung:

Finde Clusterwahrscheinlichkeiten durch Minimierung des approximierten a posteriori erwarteten Verlusts (bzgl. der Kullback-Leibler Verlustfunktion) über alle möglichen Permutationen der Clusterlabels (Stephens 2000)

Intercept/Random Effects im LCREM

- Invarianz des Intercepts und der Random Effects bzgl. einer additiven Konstanten c:

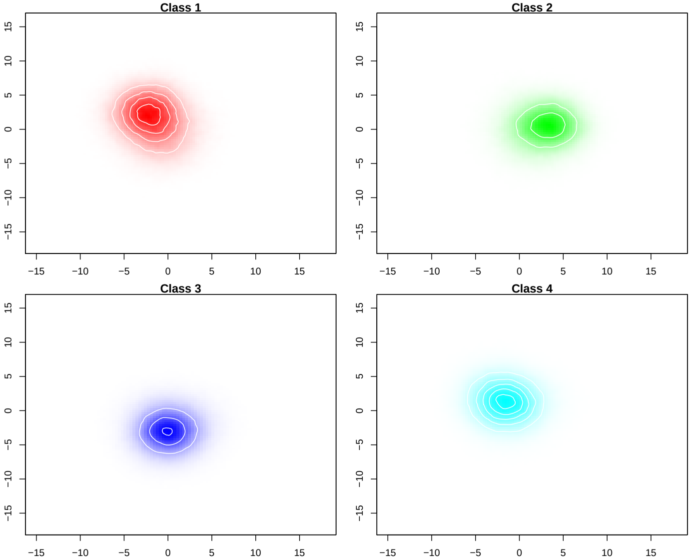
$$\delta_i^* = \delta_i + c \ , \ \forall i$$

$$\beta_0^* = \beta_0 - c$$

Lösung:

Korrelierte Vorschlagsdichten und Block-Updates im MCMC Algorithmus des LCREM (Krivitsky et al. 2009)

Vielen Dank für die Aufmerksamkeit!



Quellen I

Handcock, Mark S., Adrian E. Raftery, and Jeremy M. Tantrum. 2007. "Model-Based Clustering for Social Networks." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170 (2): 301–54.

Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. 2002. "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association* 97 (460): 1090–8.

Krivitsky, Pavel N., and Mark S. Handcock. 2008. "Fitting Position Latent Cluster Models for Social Networks with Latentnet." *Journal of Statistical Software* 24 (5): 1–23.

Krivitsky, Pavel N., Mark S. Handcock, Adrian E. Raftery, and Peter D. Hoff. 2009. "Representing Degree Distributions, Clustering, and Homophily in Social Networks with Latent Cluster Random Effects Models." *Social*

Quellen II

Networks 31 (3): 204–13.

Shortreed, Susan, Mark S. Handcock, and Peter D. Hoff. 2006. "Positional Estimation Within a Latent Space Model for Networks." *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 2 (1): 24.

Stephens, Matthew. 2000. "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62 (4): 795–809.