Ludwig-Maximilians-Universität München

Institut für Statistik

Projekt im Rahmen des statistischen Consultings

Internationaler Waffenhandel: Die Anwendung neuer Verfahren der statistischen Netzwerkanalyse

Eine Netzwerkanalyse des internationalen Kleinwaffenhandels 1992 - 2011
Kooperation mit dem Lehrstuhl für empirische Politikforschung

---

Autoren:

Roman Dieterle
roman.dieterle@hotmail.de

Felix Loewe
felixloewe@gmail.com

Projektpartner:

Prof. Dr. Paul W. Thurner

Betreuer:

Prof. Dr. Göran Kauermann

Abstract

Dieser Bericht behandelt die Analyse der *NISAT database of transfers of small arms, light weapons, and their ammunition, parts and accessories*. Die Netzwerkdaten stellen das internationale Kleinwaffenhandelsnetzwerk im Zeitraum 1992 bis 2011 dar.

Nachdem die Datengrundlage besprochen wird, erfolgt eine deskriptive Analyse des Handelsnetzwerkes anhand Zeitreihen von Netzwerkstatistiken. Im zweiten Teil wird der Querschnitt des Netzwerkes Jahr für Jahr anhand von ERGMs modelliert, um charakteristische Strukturen des Netzwerkes aufzudecken. Der Fokus liegt hierbei auf der Selektion interner Netzwerkstatistiken sowie externer Knotencharakteristika. Da dynamische Netzwerkdaten vorliegen, erfolgt im dritten Teil eine Analyse der Netzwerke anhand sogenannter *Separable Temporal Exponential Graph Models* (STERGMs, Krivitsky and Handcock, 2010). Diese Modellklasse separiert zwischen Effekten zur Tie-Formation und Effekten zur Tie-Auflösung. Die Ergebnisse werden zusammengefasst.

Inhaltsverzeichnis

1 Einführung

Die funktionale Datenanalyse beschäftigt sich mit Daten, die eine Ansammlung an Kurven darstellen.

1.1 Historisches

Die Analyse funktionaler Daten ist historisch betrachtet auf zwei Wege einzuordnen: Erstens, kann man, und das ist die elegante, statistische Klassifizierungsweise, das neue Sachgebiet im Hinblick auf die auftretenden Daten (der Stichprobenraum χ) und auf die interessierenden Größen (der Parameterraum Θ) analysieren. Die zweite Herangehensweise ist die eines angewandten Statistikers oder die von Auftraggebern, die mit einem neuen Datenformat konfrontiert sind. Sie stellen die Frage, welche historische Entwicklung dazu führt, dass ein neues Datenformat auftritt, welches neue Analysemethoden benötigt.

Cuevas (2014) wählt die erste Herangehensweise und ordnet die funktionale Datenanalyse mit Ursprung in den 1990er Jahren ein, wobei der Stichprobenraum χ ein Funktionenraum ist, und der Parameterraum Θ entweder multivariat ($\mathbb{R}^k$) oder ebenfalls ein Funktionenraum ist. Goldsmith et al. (2011) sehen den Ursprung funktionaler Daten im technologischen Fortschritt, da Messungen mit technischen Geräten genauer und hochauflösend sind. Die Messung stellt somit einen Funktionenverlauf oder ein Bild dar. Ein Bild kann als 2-dimensionale Funktion aufgefasst werden.

Beispiele für Gebiete, in denen funktionale Daten auftreten, sind Wellenmessung (Vokalkurven), Spektrometrie (Messung der Lichtabsorption von Stoffen) und Zeitreihen.

1.2 Datendesign

Die Notation der folgenden Kapitel orientiert sich an Cuevas (2014). Funktionale Daten sind eine Stichprobe zufälliger Funktionen $x_1(t), ..., x_n(t)$, die üblicherweise auf dem Intervall $[0, 1]$ standardisiert werden, sodass $t \in [0, 1]$. Sind die funktionswertigen Zufallsvariablen gemeint, so schreibt man $X_i(t)$. Die Zufallsvariablen werden in der Modellschreibweise verwendet. Das Beobachtungsintervall kann die Zeit sein oder zum Beispiel die Wellenlänge bei der Messung von Kurven. Die Funktion kann nur an einer endlichen, diskreten Anzahl von Gitterpunkten beobachtet werden. Die Gitterpunkte werden mit $t_1, ..., t_N$ bezeichnet.

2 Das lineare funktionale Modell

Zunächst wird das lineare funktionale Modell betrachtet. In diesem Modell wird *eine* skalare Zielgröße Y_i durch *eine* funktionale Einflussgröße $X_i(t)$ erklärt. Das Modell lautet (vgl. Cuevas 2014, S.10):

$$Y_i|X_i(t) \sim N(0, \sigma_i), \tag{2.1}$$

$$Y_i = \alpha + \int_0^1 \beta(t)X_i(t)dt + \epsilon_i, \quad i = 1, ..., n. \tag{2.2}$$

$\beta(t)$ bezeichnet die Koeffizientenfunktion, die die einzelnen Bereiche der funktionalen Kovariable gewichtet. Im Vergleich zu linearer Regression liegt kein Koeffizientenvektor, sondern eine Funktion vor. Außerdem werden unabhängige und identisch verteilte Fehler mit Erwartungswert 0 angenommen.

2.1 Interpretation der Gewichtsfunktion

Wie ist die Gewichtsfunktion zu interpretieren? $\beta(t)$ ist eine Funktion, die auf dem gleichen Wertebereich wie die funktionale Kovariable definiert ist. Sie soll glatt geschätzt werden. Die Interpretation lautet: Bereiche der Kovariable, in denen die Gewichtsfunktion positiv ist, haben einen positiven Einfluss auf den Outcome und vice versa für ein negatives Vorzeichen der Gewichtsfunktion. Wird $\beta(t) = 0$, haben diese Bereiche der Kovariable wenig Einfluss auf den Outcome.

2.2 Diskrete Regression

In manchen Fällen kann es ausreichen, diskrete Regressionsmodelle zu rechnen, die auf eine Glattheit der $\beta(t)$-Funktion verzichten.

Eine erste Idee ist es, eine Regression der skalaren Zielgröße auf die diskreten Gitterpunkte $t_1, ..., t_N$ durchzuführen (Ramsay und Silverman, 2005, S. 262). Es wird jeder Gitterpunkt als einzelne Kovariable betrachtet. Das funktionale Modell (2.2) wird ein multiples Regressionsmodell:

$$Y_i = \alpha + \sum_{j=1}^{N} \beta(t_j)X_i(t_j) + \epsilon_i, \quad i = 1, ..., n,$$

Da üblicherweise mehr Sampling-Punkte als Beobachtungen vorliegen, ist das Modell nicht identifiziert ($N > n$). Eine Lösung für den Koeffizientenvektor ist meist nicht intepretierbar.

Goldsmith et al. (2014, S. 832) betrachten Treppenfunktionen für $\beta(t)$. Das Intervall $[0, 1]$ wird in G Teile partitioniert: $T_1, ..., T_G$. In einem Teilintervall wird angenommen, dass die Koeffizientenfunktion konstant ist: $\beta(t) = \beta_j$ für $t \in T_j$. Für dieses Teilintervall wird der Mittelwert der funktionalen Kovariable als Regressor in einer Regression verwendet:

$$Y_i = \alpha + \sum_{j=1}^{G} \beta_j \bar{X}_{ij} + \epsilon_i, \quad i = 1, ..., n,$$

$\bar{X}_{ij}$ ist das Mittel von $X_i(t)$ auf T_j. Wird die Partition feiner gemacht, nähert sich die Treppenfunktion dem Integral an.

Eine Strategie für eine möglicherweise recht gut interpretierbare $\beta(t)$-Funktion ist es, nicht signifikante β_j auf 0 zu setzen und das Ergebnis als Treppenfunktion zu plotten. Dieser *Bin-mean approach* liefert in einer Multiplen-Sklerose-Studie zwei trennbare Gruppen, die positiv und negativ prädiktiv für den binären Outcome MS-Status sind (vgl. Goldsmith et al., 2011, S. 847).

3 Schätzung des linearen funktionalen Modells

Im folgenden wird der Schätzansatz des linearen funktionalen Modells aus Ramsay und Silverman (2005, S. 264ff.) besprochen. Die Autoren besprechen zwei Möglichkeiten: eine Schätzung mit Least-Squares, und eine Schätzung mit Penalisierungsparameter λ. Beiden gemeinsam ist, dass die Kovariablenfunktionen und die Gewichtsfunktionen mit Basisfunktionen geglättet werden.

3.1 Schätzung mit Least-Squares

Im ersten Schritt werden die Kovariablenfunktionen $X_i(t)$ werden mit einer Anzahl von K_x Basisfunktionen geglättet. Die Menge der Kovariablenbasisfunktionen wird mit $\boldsymbol{\psi}(t) = \{\psi_1(t), ..., \psi_{K_x}(t)\}$ bezeichnet. Alle geglätteten Kovariablenfunktionen, $X_i(t), i = 1, ..., n$, sind Matrixschreibweise darstellbar als

$$\mathbf{C}\boldsymbol{\psi}(t),$$

wobei $\mathbf{C}$ die $n \times K_x$-Koeffizientenmatrix bezeichnet. K_x wird üblicherweise groß gewählt, um möglichst viel der Kovariableninformation zu erhalten.

Die Gewichtsfunktion $\beta(t)$ wird mit einer Anzahl von K_β Basisfunktionen geglättet. β ist mit der Basismenge $\boldsymbol{\theta}(t) = \{\theta_1(t), ..., \theta_{K_\beta}(t)\}$ als

$$\boldsymbol{\theta}'(t)\mathbf{b}$$

darstellbar. $\mathbf{b} = (b_1, ..., b_{K_\beta})$ bezeichnet den Koeffizientenvektor der Gewichtsbasis. Das lineare funktionale Modell (2.2) kann dann in Matrixschreibweise dargestellt werden als

$$\mathbf{Y}|\mathbf{X}(\mathbf{t}) \sim N((0), \boldsymbol{\Sigma}), \tag{3.1}$$

$$\mathbf{y} = \alpha + \int_0^1 \mathbf{C}\boldsymbol{\psi}(t)\boldsymbol{\theta}'(t)\mathbf{b}\, dt + \boldsymbol{\epsilon} = \alpha + \mathbf{C}\mathbf{J}_{\psi\theta}\mathbf{b} + \boldsymbol{\epsilon}. \tag{3.2}$$

$\mathbf{J}_{\psi\theta} = \int_0^1 \boldsymbol{\psi}(t)\boldsymbol{\theta}'(t)\, dt$ bezeichnet die äußere Produktmatrix der Basisfunktionen. Sie beinhaltet das bestimmte Integral aller möglichen Produkte an Basisfunktionen der beiden Basismengen.

Dieses Modell kann per Least-Squares-Schätzung geschätzt werden:

$$\texttt{SSE}(\mathbf{y}|\alpha, \mathbf{b}) = ||\mathbf{y} - \alpha - \mathbf{C}\mathbf{J}_{\psi\theta}\mathbf{b})||^2 \rightarrow \min_{\alpha,\mathbf{b}}, \tag{3.3}$$

wobei $||\cdot||$ die Vektornorm bezeichnet. Regularisierung von β findet statt, indem K_β kleiner als K_x gewählt wird.

Ramsay und Silverman (2005, S. 275) wählen für K_β zum Beispiel den Wert 3 oder 4. Die LS-Schätzung ist weniger flexibel als die Schätzung mit Penalisierungsparameter, da K_β nur in ganzzahligen Schritten variiert werden kann (vgl. ebd.).[1]

3.2 Schätzung mit Penalisierung

Der Penalisierungsansatz bestraft große Schwankungen in β. Um große Schwankungen in der Gewichtsfunktion zu messen, wird die *Krümmung* der Kurve betrachtet. Die Krümmung einer Kurve $\beta(t)$ wird durch deren zweite Ableitung gemessen.

Ramsay und Silverman (2005) verwenden den Differentialoperator D^m, wobei m den Grad der Ableitung beschreibt. $D^2\beta(t)$ ist die zweite Ableitung der Gewichtsfunktion. Ein Maß für die Variabilität der Gewichtsfunktion ist das Integral der quadrierten zweiten Ableitung:

$$\texttt{PEN}_2(\beta) = \int (D^2\beta(t))^2\, dt$$

Durch das Quadrieren gehen negative Krümmungsbereiche und positive Krümmungsbereiche gleichermaßen in das Penalisierungskriterium mit ein.

Da $\beta(t)$ im Schätzansatz mit Basisfunktionen expandiert wird, lautet das Penalisierungskriterium (vgl. Ramsay und Silverman, 2005, S.86f.)

$$\begin{aligned}
\texttt{PEN}_2(\beta) &= \int (D^2\beta(t))^2\, dt \\
&= \int (D^2\boldsymbol{\theta}'(t)\mathbf{b})'(D^2\boldsymbol{\theta}'(t)\mathbf{b})\, dt \\
&= \mathbf{b}' \int (D^2\boldsymbol{\theta}'(t))'(D^2\boldsymbol{\theta}'(t))\, dt\, \mathbf{b} \\
&= \mathbf{b}'\mathbf{R}\mathbf{b}.
\end{aligned}$$

$\mathbf{R}$ bezeichnet die $K_\beta \times K_\beta$ Penalisierungsmatrix. Sie wird üblicherweise numerisch approximiert (z.B. Gauß-Quadratur). Wird eine B-Spline oder Fourier-Basis benutzt, ist eine analytische Berechnung möglich (vgl. Ramsay und Silverman, 2005, S.88).

$\texttt{PEN}_2(\beta)$ wird als Penalisierungskriterium verwendet, indem es zur Residuenquadratsumme (3.3) addiert wird. Es entsteht die penalisierte Residuenquadratsumme

$$\texttt{PENSSE}_\lambda(\mathbf{y}|\alpha, \mathbf{b}) = ||\mathbf{y} - \alpha - \mathbf{C}\mathbf{J}_{\psi\theta}\mathbf{b}||^2 + \lambda\mathbf{b}'\mathbf{R}\mathbf{b}, \tag{3.4}$$

die nach α und $\mathbf{b}$ minimiert wird. Der Glättungsparameter λ steuert die Glattheit der Gewichtsfunktion. Für $\lambda \to 0$ wird keine Rauheit bestraft. Für $\lambda \to \infty$ muss $\beta(t)$ eine lineare Funktion ohne Krümmung sein.

[1] Die Least-Squares-Schätzung mit B-Splines als Basisfunktionen wird in der Literatur auch „regression spline smoothing“ genannt. Bei der Schätzung mit Penalisierung sprechen Ramsay und Silverman von „spline smoothing“.

4 Das generalisierte funktionale Modell

Im folgenden wird das generalisierte funktionale Modell vorgestellt [GBC+11], welches eine Erweiterung des linearen funktionalen Modelles (2.2) darstellt. Dieses Modell ist ein gemischtes Modell, in welchem eine skalare Zielgröße Y_i durch eine funktionale Einflussgröße $X_i(t)$ und mehreren skalare Kovariablen Z_i erklärt wird.

5 Beispiel mit Spektrometrie-Kurven

6 Zusammenfassung

Diese Arbeit

Literaturverzeichnis

[Cue14] Cuevas, A. (2014): A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147, 1-23.

[GBC+11] Goldsmith, J., Bobb, J., Crainiceanu, C.M., Caffo, B.S., Reich, D.S. (2011). Penalized Functional Regression. *Journal of Computational and Graphical Statistics 20*, 830-851.

[RS05] Ramsay, J.O. and Silverman, B.W. (2005). Functional Data Analysis. New York: Springer.

[R14] R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich diese Arbeit ohne fremde Hilfe angefertigt und nur die im Literaturverzeichnis aufgeführten Quellen und Hilfsmittel benutzt habe. Diese Arbeit wurde noch nicht zu anderen prüfungsrelevanten Zwecken vorgelegt.

.....	.....
Ort, Datum	*Roman Dieterle*